Argument Quality Assessment in the Age of Instruction-Following Large Language Models

Henning Wachsmuth,[1] **Gabriella Lapesa,**[2] **Elena Cabrio,**[3] **Anne Lauscher,**[4]
Joonsuk Park,[5] **Eva Maria Vecchi,**[6] **Serena Villata,**[3] **Timon Ziegenbein**[1]
[1]Leibniz University Hannover, {h.wachsmuth,t.ziegenbein}@ai.uni-hannover.de
[2]GESIS, Heinrich-Heine University Düsseldorf, gabriella.lapesa@gesis.org
[3]Université Côte d'Azur, CNRS, Inria, I3S {elena.cabrio,serena.villata}@univ-cotedazur.fr
[4]Universität Hamburg, anne.lauscher@uni-hamburg.de
[5]University of Richmond, park@joonsuk.org
[6]University of Stuttgart, eva-maria.vecchi@ims.uni-stuttgart.de

Abstract

The computational treatment of arguments on controversial issues has been subject to extensive NLP research, due to its envisioned impact on opinion formation, decision making, writing education, and the like. A critical task in any such application is the assessment of an argument's quality—but it is also particularly challenging. In this position paper, we start from a brief survey of argument quality research, where we identify the diversity of quality notions and the subjectiveness of their perception as the main hurdles towards substantial progress on argument quality assessment. We argue that the capabilities of instruction-following large language models (LLMs) to leverage knowledge across contexts enable a much more reliable assessment. Rather than just fine-tuning LLMs towards leaderboard chasing on assessment tasks, they need to be instructed systematically with argumentation theories and scenarios as well as with ways to solve argument-related problems. We discuss the real-world opportunities and ethical issues emerging thereby.

Keywords: Computational Argumentation, Argument Quality, Large Language Model, Instruction Fine-Tuning

1. Introduction

"In some sense, the question about the quality of an argument is the 'ultimate' one for argumentation mining." (Stede and Schneider, 2018).

When learning about controversial issues, people rarely accept arguments they encounter without further contemplation. Rather, they seek to find *the best* arguments; those that help them form an opinion or write texts that persuade others; those that make them reach agreement or at least understand each other better. That is to say, *argument quality* is of interest as soon as arguments are presented to an audience. Computational argumentation aids the treatment of arguments at a larger scale, with important applications in search (Wachsmuth et al., 2017c), business (Slonim et al., 2021), and education (Wambsganss and Niklaus, 2022). But the situation there is the same: It is not enough to mine or generate arguments; their quality also needs to be evaluable (Park and Cardie, 2018), so that it can be assessed (Lauscher et al., 2020), flaws can be found (Goffredo et al., 2022), and accounted for (Skitalinskaya et al., 2023).

Wachsmuth et al. (2017b) surveyed research on argument quality assessment, organizing theories and methods under 15 quality notions, from logical cogency to rhetorical effectiveness to dialectical reasonableness. Even though computational argumentation was just gaining momentum in natural language processing (NLP) back then, rarely going beyond argument mining, two inherent challenges of argument quality were visible already: the *diversity* of quality notions as well as the *subjectivity* of their perception and, hence, of their assessment for both humans and computational models. Consider the following argumentative claim against censoring Mark Twain's usage of the N-word, taken from the debate platform kialo.com:

"In Huckleberry Finn, Twain captured the essence of everyday midwest American English."

This claim is certainly relevant to the discussion, but whether people will deem it effective may strongly depend on their individual context. A person without African-American background may be willing to accept the argument; one with high literacy might look for clearer logical connections.

While the challenges of diversity and subjectivity prevail until today (Lapesa et al., 2023), NLP is now seeing a revolutionary breakthrough: the rise of *instruction-following large language models* (henceforth, LLMs) that can tackle various NLP tasks with little to no task-specific fine-tuning, enabled by their supreme capability to integrate and leverage knowledge across contexts (OpenAI, 2023). The question is: *What are the implications for argument quality assessment specifically as well as for computational argumentation in general?*

In this position paper, we revisit the computational assessment of argument quality in light of the availability of LLMs such as GPT-4 and Alpaca

(Taori et al., 2023). Starting from the status quo reported by Wachsmuth et al. (2017b), we carry out a brief survey of recent NLP research on the topic (Section 2). To bring order into the various lines of research pursued since 2017, we organize them into three general directions, as laid out in Figure 1:

- *Conceptual notions* of maximal and minimal argument quality,
- *Influence factors* of argument quality from the context where arguments occur, and
- *Computational models* for assessing or improving argument quality.

On this basis, we establish the central question to which we provide answers in this paper:

> *How to drive research on LLM-based argument quality assessment in order to face the prevailing challenges of diverse quality notions and their subjectivity?*

In particular, we are convinced that the capabilities of instruction-following LLMs enable research to overcome many aspects of the two challenges. To this end, the primary focus of NLP research on argument quality should be put on systematic ways to teach LLMs to follow instructions, including concepts and settings of arguing in addition to ways to solve argument-related problems (Section 3). Instead of fine-tuning LLMs on predefined domains (manifested in the training data) and preselected theories (manifested in the data's annotations), as well as simple engineering of prompts, we expect the greatest impact to lie in teaching LLMs the theories, circumstances, and ethical constraints to adhere to. The rationale behind this is that LLMs will often have processed data from all contexts needed to make an informed judgment about an argument's quality, due to their heavy pretraining on huge amounts of data. In contrast, LLMs cannot access, by default, the knowledge of what is to be prioritized in a given setting.

We state upfront that the blueprint delineated in this paper comes with several limitations and ethical considerations that we critically analyze below. Moreover, we are naturally aware of the general issues of LLMs, including hallucinated facts and the reproduction of common social biases. These issues deserve treatment in computational argumentation as well; they are even particularly critical there due to the sensitivity of many controversial topics (Holtermann et al., 2022). Keeping this in mind, we believe that it is necessary to explore now how to best employ LLMs for argument quality assessment in order to harness their full potential for the main applications, while avoiding to waste energy for the typical pursuit of leaderboard rankings on existing quality assessment tasks. This is the goal of the paper at hand.

Now, why is it important to discuss LLMs for argument quality assessment specifically? We address this matter when we look at the real-world opportunities emerging from the capabilities of LLMs in academia and industry (Section 4). While Argyle et al. (2023) developed LLMs that tone down argumentative conversations, we postulate a contrary path: Exploiting the means of LLMs to proactively enable people to learn and better reason about controversial issues, thus contributing towards more deliberate conversations (Vecchi et al., 2021). We think that the time has come to revisit and pursue the core visions of computational argumentation research, from the overcoming of filter bubbles to the individualized mass education of learners. We sketch how these visions could be realized with the LLMs available today, before we conclude (Section 5) and stress ethical concerns that arise with LLMs that actively affect human views (Section 6).

With the discussion in this position paper, we provide two main contributions to research:

1. *A survey* of the main lines of recent research on argument quality and its assessment
2. *A blueprint* for impactful future research on LLMs for argument quality assessment

2. A Brief Survey of Recent Research

To start, this section briefly but systematically surveys recent NLP work on argument quality assessment. We identify three general directions, each with two main aspects, and organize the research accordingly, as illustrated in Figure 1.

2.1. Frame of the Survey

Beyond holistic computational argumentation (CA) surveys (Stede and Schneider, 2018; Cabrio and Villata, 2018; Lawrence and Reed, 2019), Wang et al. (2023a) specifically reviewed works on argument generation, Lauscher et al. (2022b) on knowledge in CA, and Vecchi et al. (2021) on the use of CA for the social good. Also, some tutorials treated argument mining and its applications (Budzynska and Reed, 2019; Bar-Haim et al., 2021). In contrast, we focus on argument quality assessment.

Aside from our recent tutorial (Lapesa et al., 2023), the only argument quality review that we are aware of is the one of Wachsmuth et al. (2017b) who organize relevant literature until mid 2017 into a taxonomy of 15 logical, rhetorical, and dialectical quality dimensions. The authors already discussed the diversity of quality notions as well as the subjectivity of their perception, both of which are still hampering research. In this paper, we seek to delineate ways to overcome them, in line with the authors' organization of what argument quality means. We

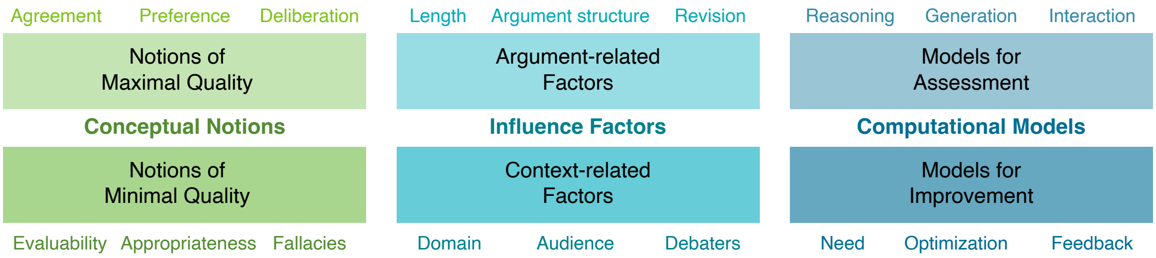

Figure 1: Organization of the surveyed argument quality research into three general directions (conceptual notions, influence factors, and computational models), their main aspects (e.g., notions of maximal and minimal quality), and specific concepts studied for these (e.g., agreement, preference, and deliberation).

start from their work here, so we restrict our survey to work that is published after theirs.

Based on our experience with NLP research on CA, we cover four groups of publication venues:

- All NLP venues covered by the ACL anthology
- The leading artificial intelligence (AI) conferences, all from AAAI.org and IJCAI
- The leading information retrieval (IR) conferences, SIGIR and ECIR
- The leading CA conference, COMMA

We used Google site search and Springer and ACM's internal search on September 1, 2023 (updated on October 17, 2023), to gather all papers containing any of the following pairs of words:

$$\begin{aligned} &\{\textit{argument, argumentation, debate}\} \\ \times\ &\{\textit{quality, strength, persuasiveness}\} \end{aligned}$$

This led to 257 papers (202 NLP, 35 AI, 11 IR, 9 CA). From these, we kept all 119 papers that deal with quality of natural language arguments based on title, abstract, and skimming (98 NLP, 12 AI, 6 IR, 3 CA). To focus on scientific novelty, we further excluded surveys, tutorials, demos, shared tasks, and system papers, leaving 104 papers (87 NLP, 10 AI, 5 IR, 2 CA). We checked these in more detail to ensure that argument quality is actually part of the research. After filtering out others, we obtained a final set of 83 papers (69 NLP, 9 AI, 3 IR, 2 CA).

2.2. General Research Directions

Analyzing the 83 papers as a whole, we identified the following three general directions of research on argument quality, each with two main aspects. Concretely, one author of this paper proposed the organization. The papers were then distributed among all other authors who ranked them by the directions and aspects they contribute, if any.

Conceptual Notions 24 of the papers primarily deal with the question of what is actually meant by argument quality, considered from either of two complementary perspectives:

- *Notions of maximal quality* based on arguing goals such as agreement and deliberation or on preferences between different arguments
- *Notions of minimal quality* in terms of what makes an argument evaluable or appropriate to be stated as well as how to avoid fallacies

Influence Factors This direction covers 30 papers studying (or controlling) two types of factors that influence the perception of quality beyond the content, structure, and style of the argument itself:

- *Argument-related factors* such as the argument's length, its structure in terms of relations between units, and revisions applied to it
- *Context-related factors*, such as the domain of the discussion, the audience addressed, and the debaters involved

Computational Models Finally, 21 papers aim mainly at methodological novelty in the modeling of argument quality for two quality-related tasks:

- *Models for assessment* of argument quality, capturing specificities of the task, the whole discussion, or the context of arguing
- *Models for improvement* of argument quality, targeting the need for improvement, actual optimizations, or feedback on what to improve

The remaining eight papers pursue individual research directions. We note that many of the surveyed papers do not fall under one general direction only; rather, they often have a visible focus on one of them. In particular, our internal discussion revealed that contributions to influence factors and computational models are not always easy to distinguish and that, sometimes, models may rather target downstream applications. Still, the directions and aspects were agreed upon in general.[1]

[1] For validation, we reassigned 16 papers (19%) to other authors: 11 got the same main direction; for four, it was seen as the second contribution. Only in one case, a fully different direction was assigned. After rechecking, the newly assigned direction did not seem adequate.

In the following, we discuss selected works from each of the general research directions. Table 1 in the appendix shows the full list of all 83 covered publications, grouped by the primary general research direction and the main aspect.

2.3. Conceptual Notions

Naturally, all surveyed literature builds on some notion of argument quality, at least implicitly. However, we found that 24 of the 83 papers have the explicit treatment of quality notions as their main focus and 10 further papers contribute to quality notions to some extent. About two-thirds of the works discuss how an argument should be ideally (maximal quality), the others what an argument should at least achieve or avoid (minimal quality).

Notions of Maximal Quality Some researchers build on the argument quality taxonomy proposed by Wachsmuth et al. (2017b), including Lauscher et al. (2020) who model the main taxonomy notions using multitask learning across Q&A, debate, and review forums. Others question the simplifying view that argument quality is about persuasion only: El Baff et al. (2018) consider the goal of *agreement*, defining good news arguments as those that challenge or corroborate stance. Gretz et al. (2020) see argument quality as a *preference* relation, and Falk et al. (2021) examine its connection to *deliberation*. With an entirely different perspective, some papers examine what makes an argument good irrespective of topic (Beigman Klebanov et al., 2017), whereas, for example, Dumani et al. (2020) operationalize argument quality for practice in a quality-based framework for argument retrieval.

Notions of Minimal Quality Park et al. (2015) establish the notion of an argument's *evaluability*, that is, the prerequisite of assessing logical quality soundly. A key research line on minimal quality is the detection of *fallacies*: arguments with flawed or deceptive reasoning. Neural models have shown success on this deep semantic problem; some aim at ad-hominem arguments only (Habernal et al., 2018), others at various fallacies (Jin et al., 2022). Persing and Ng (2017b) tackle the broader problem of spotting an argument's weaknesses, from grammar errors to lack of objectivity and unclear justifications. More practice-oriented, Pauli et al. (2022) look at the misuse of fallacies for rhetorical appeals in online forums and fake news. Finally, Ziegenbein et al. (2023) refine the notion of *appropriateness*, emanating from Aristotle's work (Aristotle, ca. 350 B.C.E./ translated 2007). They see it as the minimal quality that makes arguments worthy of being considered and annotate data for violations of appropriateness.

2.4. Influence Factors

Assessing the different notions of argument quality is a complex task and is influenced by many factors, some of which have no explicit relation to the argument itself. Accordingly, research has dealt with the identification, modeling, and controlling of such factors and their impact on argument quality. We found that 30 of the 83 papers mainly focus on influence factors, and a further 19 papers are to some extent devoted to them. Of these, about 60% discuss argument-related factors while the rest looks at context-related factors.

Argument-related Factors In terms of textual factors influencing the perceived quality of arguments, researchers display the questionable power of *length* as a predictor (Potash et al., 2017) and account for this in dataset creation (Toledo et al., 2019). The impact on quality of internal *argument structure* has been investigated using the notion of organization quality in learner essays (Chen et al., 2022a) and by using annotations of argument components in business model pitches (Wambsganss and Niklaus, 2022). Notions of structure within an argument are further extended through adding attributes to argument components (Carlile et al., 2018), shifting the focus to component-related factors, or by comparing different *revisions* of the same claim (Skitalinskaya et al., 2021).

Context-related Factors Lukin et al. (2017) analyze the interaction between argumentative styles (emotional vs. factual) and the personality of the *audience*, as modeled by the Big Five traits. Similarly, Durmus and Cardie (2018) model political and religious ideologies, based on the audience's stances on various controversial topics. Both indicate that audience-level factors often outweigh language use in their persuasive effect. Alshomary et al. (2022) turn the view to rhetorical strategies of *debaters*, assessing the effect of morally-framed arguments. They find that morals are particularly successful in challenging the audience's beliefs. Wiegmann et al. (2022) analyze stylistic and behavioral characteristics of debaters that contribute to their persuasiveness over multiple debates. Aside from debate participants, Liu et al. (2022) explore arguments on social media that are accompanied by images, highlighting the potential of multimodal approaches to quality assessment, whereas Fromm et al. (2023) generalize the contextual scope of assessment to multiple *domains* at the same time.

2.5. Computational Models

The majority of the 83 surveyed papers include empirical experiments with models for argument quality. However, we found that only 21 of them actually focus on proposing novel approaches targeting either of the above-mentioned conceptual quality notions, whereas 26 other papers have such approaches as secondary contributions to support their claims with experimental results and analysis. Almost all approaches aim at the assessment of argument quality, but a few recent ones go beyond assessment, studying how to improve quality.

Models for Assessment Many approaches aim at specific quality notions. For example, the attentive interaction model of Jo et al. (2018) predicts an opinion holder's view change by detecting vulnerable regions in their *reasoning* and modeling its relation to a challenger's argument. Gleize et al. (2019) propose a Siamese neural network to assess the convincingness of evidence, while Song et al. (2020) develop a hierarchical multitask learning approach to jointly model discourse element identification and organization assessment for essay scoring. Gurcke et al. (2021) examine to what extent an argument's logical sufficiency can be predicted based on whether its conclusion can be inferred from its premises using the *generation* capabilities of transformers. Kondo et al. (2021) assess the validity of an argument's reasoning using Bayesian networks and predicate logic facilitated by argumentation schemes. A few works also look beyond single quality notions, such as Falk and Lapesa (2023) who inject knowledge about the *interactions* between different quality notions to improve the prediction of individual ones.

Models for Improvement While only a few models for improvement have been presented so far, we expect more to come soon, also seeing related efforts on topics beyond those covered in this survey (Chakrabarty et al., 2021; Ihori et al., 2022; Li et al., 2022). An early attempt was made by Ke et al. (2018) who design neural models that predict the persuasiveness and other attributes of arguments in a student essay, to provide *feedback* to students on how to improve their arguments. Recently, Skitalinskaya and Wachsmuth (2023) identified arguments in *need* of improvement, leveraging complex revision-based data with transformer models. Skitalinskaya et al. (2023) go one step further, presenting the first approach to the *optimization* of argumentative claims. It combines neural claim rewriting with quality-based ranking.

2.6. Other Research Directions

Among the eight papers that do not fit under the three main research directions, we identified the following two rough research areas.

Five papers deal with specific applications for which argument quality assessment is key. Rach et al. (2020) and Kiesel et al. (2020) target argument search, both taking a human-interaction perspective: The former studies the effects of integrating argument search into an avatar-based dialogue system; the latter investigates user expectations on voice-based argument search systems, such as preferred ranking criteria. Chalaguine and Hunter (2020) develop a chatbot that relies on an argument graph for persuasive counterargument generation, and Falk et al. (2021) address expert moderation in a deliberative forum, taking moderator interventions as implicit labels for the need to improve comment quality. Fromm et al. (2021) use argument mining for analyzing peer reviews.

The other works tackle the bottleneck of (scarce) assessment training data. Heinisch et al. (2022) employ data augmentation to support the prediction of argument validity and novelty. Kees et al. (2021) evaluate active learning strategies for supporting argument strength estimation, and Yang et al. (2019) introduce a quality control method that they apply to annotate argument acceptability.

3. LLMs for Argument Quality

Section 2 stresses that a big part of argument quality research tackles the key challenges of diverse views of quality (by developing or refining notions) and subjectivity (by controlling or modeling influence factors). However, the intricated interdependencies between different quality notions and the various factors that influence them have hampered substantial progress in the reliable assessment of argument quality so far. We argue that instruction-following large language models (LLMs) have the potential to overcome many limitations, if systematic ways to teach them accordingly are established. In this section, we start from the main advantages of such LLMs. Then, we outline what to instruct LLMs with and how to do so (beyond simple in-context learning/prompting techniques) in order to advance LLMs for argument quality in future research.

3.1. Assessment without Instructions

Conceptually, argument quality assessment is a classification or regression problem, even if partly treated as preference learning. For decades, research in NLP relied on the traditional supervised learning paradigm for most such problems: to induce a mapping from one representational space to another using training pairs of input and output. As sketched in Figure 2a, the input spaces (usually, representing natural language) and output spaces (label schemes or value ranges, such as argument quality scores) are separated thereby. This separation prevents any exchange of knowledge across spaces and across tasks.

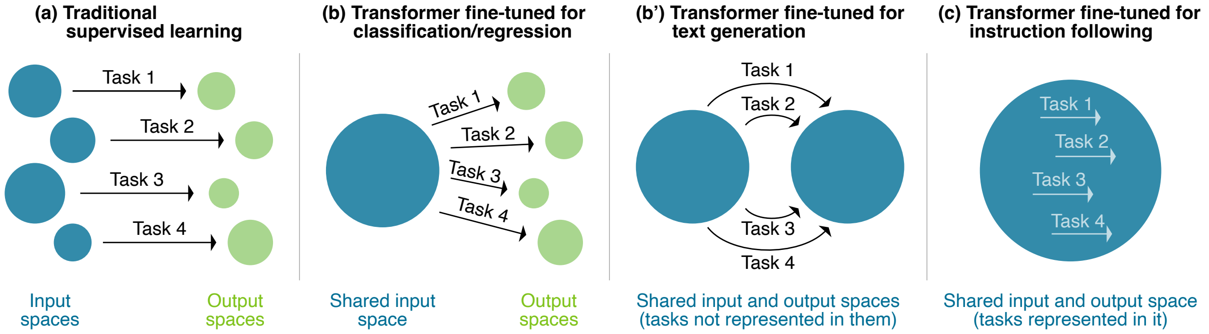
(a) Traditional supervised learning
Task 1
Task 2
Task 3
Task 4
Input spaces
Output spaces
(b) Transformer fine-tuned for classification/regression
Task 1
Task 2
Task 3
Task 4
Shared input space
Output spaces
(b') Transformer fine-tuned for text generation
Task 1
Task 2
Task 3
Task 4
Shared input and output spaces (tasks not represented in them)
(c) Transformer fine-tuned for instruction following
Task 1
Task 2
Task 3
Task 4
Shared input and output space (tasks represented in it)

Figure 2: Learning of representational spaces in NLP models (same color: same type of representation): *(a) Traditional supervised learning:* Input and output spaces are separated across tasks; representations are task-specific. *(b) Classification/Regression transformer:* The input space is shared across tasks; its representation can be learned on all tasks. *(b') Generation transformer:* Both spaces are shared across tasks; their representations can be learned on all tasks, but not task interactions. *(c) Instruction-following transformer:* One space for inputs, outputs, and tasks; representations can be learned jointly on all tasks.

With the shift to transformers in NLP, the learning effort is mostly reduced to the self-supervised pretraining of a language model (Vaswani et al., 2017; Devlin et al., 2019). Under the transfer learning paradigm, only fine-tuning remains supervised, to make a model address the task it is supposed to. This way, knowledge is shared between input representations across tasks and contexts, that is, all texts ever processed affect how an input is encoded. In classification and regression, however, fine-tuning (say, of a BERT encoder with a quality scoring head) reintroduces a key restriction of traditional methods, illustrated in Figure 2b: The input space is separated again from the output space, preventing models from fully leveraging knowledge acquired from solving other tasks.

Fine-tuning for text generation tasks keeps input and output in the same space; thereby, for example, connections between an argument and its improved version can be learned. Still, it faces a second restriction that is shown in Figure 2b': The idea behind the mapping from input to output (why is an output correct for an input) remains fully implicit in the training pairs. For argument quality assessment, both restrictions imply that only those interdependencies between quality notions as well as those contextual influence factors are taken into account that are explicitly modeled or controlled by the human developer. Even though a lot of other interdependencies and factors may be well-known in argumentation theory (van Eemeren and Grootendorst, 2004; Aristotle, ca. 350 B.C.E./ translated 2007), they are widely ignored thereby. This is where instruction-following LLMs go beyond.

3.2. Instruction-Following LLMs

Instruction fine-tuning teaches LLMs to follow user instructions to solve tasks (Peng et al., 2023). As Figure 2c stresses, it does so by representing the task in the joint space of inputs and outputs across all tasks and contexts, that is, in natural language (more precisely, in the embedding space in which language is encoded and from which it is decoded). Following the instruction fine-tuning paradigm, deriving an output from an input remains a language modeling problem as well as how to operationalize the derivation. This means that all knowledge ever processed in pretraining is still accessible in principle (bounded by the technical constraints of the model). Then, the supreme capability of transformers to integrate and leverage knowledge across contexts enables instruction-following LLMs to tackle unseen tasks with no or little fine-tuning.

In argument quality assessment, we can expect that most knowledge about quality notions and their interdependencies as well as about influence factors and their effect on the subjective perception of argument quality has already been processed by leading LLMs, such as GPT-4 (OpenAI, 2023) and Alpaca (Taori et al., 2023), in their pretraining stage. It should thus be possible to learn through instructions what is important in the assessment task at hand while not ignoring interactions with the surrounding concepts of argument quality.

3.3. Instructions for Assessment

Fine-tuning LLMs on general-purpose instruction data (Wang et al., 2023b) will help them solve language modeling tasks in principle. By default, however, LLMs do not necessarily have access to what is to be prioritized for the setting of the task at hand. We expect that this is the information an LLM needs to be instructed with to assess argument quality reliably. Accordingly, we see the survey results from Section 2 as an adequate basis to explore what to teach LLMs for the assessment. Instructions may thus include but are not limited to:

- arguing goals, from agreement (El Baff et al., 2018) to deliberation (Falk et al., 2021);
- definition of various quality notions, be it for maximal quality (Lauscher et al., 2020) or minimal quality (Jin et al., 2022);
- specificities of audiences (Lukin et al., 2017) and debaters (Durmus and Cardie, 2019);
- background on controversial topics, such as other arguments (Luu et al., 2019) or relationships between the topics (Zhao et al., 2021);
- ethical aspects, such as biases (Holtermann et al., 2022) and culture (Chen et al., 2022a);
- any examples of respective assessments, following the common few-shot learning idea.

Exemplarily, let us get back to the Huckleberry Finn claim from Section 1 taken from kialo.com. On this online platform, users create and refine claims, and they give impact votes from 1 to 5. These impact votes may serve as gold labels. For voting, users are asked to consider both a claim's persuasiveness and its relevance to the claim it replies to, in equal weight.[2] When using such labels, supervised learning (Figure 2a) and standard transformer fine-tuning (Figure 2b/b') can learn the semantics of the task only through the mapping from claim to vote, risking spurious correlations (Thorn Jakobsen et al., 2021) as well as bias (Spliethöver and Wachsmuth, 2020). The equal weighting will likely not to be captured either, as users may not weigh systematically, may not have read instructions, or may just have a subjective perception.

Moreover, the same claim should certainly be assessed differently depending on the setting (beyond kialo.com). From a deliberative perspective, for example, it lacks a concrete counterproposal (e.g., *even if we do not censor, a preface should be added or teachers should discuss the load of the N-word in everyday life*). Moving to quality improvement, the claim may need to be revised for specific audiences, for example, the concept of "everyday midwest American English" may be completely opaque to people with low literacy.

Instructions may overcome all these challenges. An example may be: *"Rate the claim's quality from the perspective of deliberation, when presented to a person of low literacy"*. This makes the semantics of the task much more explicit, likely reducing biases and spurious correlations. It performs a stage-setting and an addressee-setting function which are crucial for assessment and improvement. We do not aim to come up with the best instructions in this paper, but see this as a task for future work. Rather, we discuss how to systematize respective instruction fine-tuning attempts.

3.4. Blueprint for Instruction Fine-Tuning

Effective processes for the general instruction fine-tuning of LLMs have already been established in prior work (Taori et al., 2023). Given the discussed advantages of such LLMs over previous models, we argue that argument quality assessment may be brought to the next level through systematic approaches to *task-related* instruction fine-tuning. The idea is to bring specific knowledge about theories, circumstances, and ethical constraints of arguing along with ways of how to solve argument-related problems into the fine-tuning process. As a blueprint, such an approach could roughly consist of the following stages:

1. Start from a general instruction-following LLM, such as Alpaca (Taori et al., 2023). Even some standard pretrained transformer may suffice, if general instructions are added to Step 2.
2. Acquire a seed set of argumentation-specific instructions, covering concepts such as those in the previous subsection. For example, these instructions can be derived manually or semi-automatically from the various datasets and experiments covered in the surveyed papers.
3. Depending on available resources, apply techniques such as reinforcement learning using human feedback (Ouyang et al., 2022), fine-tuning on self-generated instructions (Wang et al., 2023b), or other instruction fine-tuning mechanisms that are proposed in research.
4. Align the behavior of the instruction fine-tuned LLM on new unseen tasks at hand using systematic prompt design, for example, via soft prompting (Qin and Eisner, 2021) or sociodemographic prompting to emulate social profiles of debaters and audiences (Beck et al., 2023). Due to Step 3, these tasks now benefit from argument-specific task-solving skills.

At least for fact-related argument quality dimensions, such *local acceptability* (Wachsmuth et al., 2017b), an additional step may be to systematically work against hallucinations, by teaching the LLM to check arguments against some fact source (e.g., a knowledge base or a corpus). Factuality measures may be included in the model optimation for this purpose. We note, though, that this presupposes a setting in which sources can also be accessed at inference time. Besides, many quality dimensions are actually not (inherently) about facts, such as those from rhetorics (Wachsmuth et al., 2017b).

We expect that the resulting LLM will assess argument quality more reliably in line with the theories behind diverse quality notions and will adjust to the subjective viewpoints of interest. Thereby, various new opportunities emerge for real-world applications, as discussed in the next section.

[2] https://support.kialo-edu.com/en/hc/about-voting/

3.5. Evaluation of Quality Assessment

Finally, we make a note on the evaluation of LLM-based argument quality assessment, to give a basic guideline. Various ways of evaluating assessment have been pursued in prior work; particularly, there is a debate about whether quality should be assessed in absolute terms, based on a given score range, or in relative terms, comparing different arguments to one another (Wachsmuth et al., 2017a). Instruction-following LLMs might not entirely resolve the issue behind; while they provide new means for a reliable evaluation (e.g., handling of context), their generative nature may also complicate the validation against some ground-truth.

Ultimately, a fully *unified* evaluation procedure may not be possible, because it depends on what information is available in a given assessment setting. Rather, we propose that an evaluation procedure *ideally* takes into account the main decisive factors of quality, as we exemplified for the Huckleberry Finn claim above: What quality dimension is of interest, who is the audience of the argument, and similar. The evaluation may happen on a careful selection of existing datasets, but new benchmarks that account for the factors may be needed, too.

Criteria-wise, we see a mix of absolute and relative assessment as best approximating how humans assess quality, but this requires careful operationalization: Many argument quality dimensions imply some hard constraints, which speaks for an absolute part (e.g., are the argument's premise acceptable?). However, there may not be a clear best/worst quality for an argument, which speaks for a relative part wherever other arguments are accessible (e.g., are the premises more acceptable than those of other arguments?). Instruction fine-tuning should prepare an LLM for dealing with both parts and, hence, be evaluated against them.

4. Opportunities for the Real World

Arguably, instruction-following LLMs generally provide great opportunities for NLP and its applications. Their wide and easy applicability, along with their often low need for task-specific training data, is particularly beneficial in the context of interdisciplinary research. With a successful realization of the blueprint delineated above, however, we explicitly see specific potential for computational argumentation applications, due to their inherent need for argument quality assessment (see Section 1). We now sketch some of the main opportunities we see. Partly, they bring up ethical concerns, though, that we discuss at the end of the paper (Section 6).

Debating Technologies So far, one of the most impressive applications is IBM's Project Debater, which has competed well with professional human debaters (Slonim et al., 2021). Its quality assessment methods are audience-agnostic (Toledo et al., 2019; Gretz et al., 2020), which may not suffice to convince people across diverse backgrounds, as research indicates (Alshomary et al., 2022). Moreover, Project Debater's arguments are retrieved, recomposed, and rephrased rather than written naturally. If controlled well, the generation capabilities of LLMs may advance notably on the latter, whereas our proposed fine-tuning process may explicitly target the adjustment to audiences.

Argument Search Argument search aims to find the best pros and cons on controversial topics. Unlike for debating technologies, its goal to aid self-determined opinion formation suggests not to tune towards audiences. However, argument search engines miss a reliable quality-based ranking so far (Wachsmuth et al., 2017c; Stab et al., 2018; Dumani et al., 2020), likely due to the heterogeneity of argumentative domains and genres on the web. The low training need of instruction-following LLMs may alleviate this shortcoming. In addition, the text rewriting capabilities of LLMs may be employed to optimize the presentation of arguments (Skitalinskaya et al., 2023), or to fill gaps as needed. We expect that convincing rankings and presentations are key to making people open to argument search, enabling them to overcome filter bubbles.

Discussion Moderation The moderation of (online) content is critical to ensure healthy and productive discussions (Park et al., 2012, 2021; Vecchi et al., 2021). This holds particularly for deliberative contexts, where participants should be supported in communicating their viewpoints. Effective moderation reaches a bottleneck as the scale of online discussions grows (Klein, 2012; Shortall et al., 2022). LLMs instructed for argument quality can assist moderation efforts by detecting possible violations of community guidelines, inappropriate language, or generally low-quality arguments in discussions. This way, moderators can focus their attention on nuanced cases and appeals, optimizing efficiency and ensuring a healthier discourse. In some settings, generative LLMs could even lead a dialogue with users to provide clarifications, feedback, and improvement suggestions.

Argumentative Writing Support LLMs may further provide individualized education to learners (e.g., students or non-native speakers) as well as to everyday writers (e.g., e-commerce customers), for instance, by giving feedback on the quality of their arguments (Carlile et al., 2018; Chen et al., 2022a). Instruction fine-tuning makes it easier to go beyond simple quality scoring (e.g., how clear an argument is) to targeted hints (e.g., *Provide more evidence for your initial claim!*). Prompted with the writing goal, LLMs may also suggest argument completions, such as missing conclusions (Gurcke et al., 2021). With these means, students may learn to reason more soundly, product reviews can become more informative, and so forth. LLMs instructed with the concrete feedback scenario (e.g., *a student learning to write essays in English*) will help to further individualize support and may even adjust to the specific learning need of the user.

Other Applications In several other scenarios, argument quality is crucial. One example is to *generate summaries* of the best arguments in news articles (Syed et al., 2020) or online discussions (Syed et al., 2023). Here, instruction-following LLMs can interpret the term *best* as needed—without any task-specific fine-tuning. In the *medical domain*, argument quality plays a central role for evidence-based medicine (Mayer et al., 2021). A well-instructed LLM may assess evidence strength, thus enabling better inferences based on clinical trials or reports. Similarly, the reasonableness of arguments on health discussion online platforms may be evaluated. Further scenarios include *e-commerce*. There, an LLM-based service chatbot can, for example, select arguments based on quality notions (e.g., clarity) to explain to customers why a request cannot be completed, to minimize their dissatisfaction. Argument quality may also be assessed in *recommender systems* to make justified suggestions based on compelling reasons.

Implications for Research Finally, we also see great potential for diversity and subjectivity-aware instruction fine-tuning when it comes to driving fundamental research, as sketched here for two examples: interdisciplinary work at the interface of NLP and computational social science, and the methodological development driven by the need to cope with subjectivity in argument quality annotations.

The social science context adds even more *diversity*, including sophisticated quality notions and domain-specific language, along with new challenges, such as well-curated and annotated, but small and imbalanced datasets (Falk and Lapesa, 2022). Our instruction fine-tuning blueprint fits exactly such scenarios: annotation guidelines serve as instructions, highly-curated annotations as reinforcement examples, and the knowledge encoded in LLMs alleviates resource-lean issues. Additionally, the scene-setting function of instruction fine-tuning (see Section 3) has the potential to address the deliberative goal of defining and quantifying discourse quality across contexts (Esau et al., 2021).

The multiple factors of *subjectivity* influencing argument quality perception (debater and audience beliefs, values, etc.) often limit the inter-annotator agreement (Wachsmuth et al., 2017b). Ultimately, subjectivity is a constitutive feature of argument quality, as indicated above. Romberg (2022) suggest to join the perspectivist turn of machine learning and NLP (Plank, 2022; Cabitza et al., 2023) in computational argumentation. LLMs' perspective-taking capabilities could be a game changer for this, assuming that the risks of sociodemographic prompting (Beck et al., 2023) and stereotypes (Cheng et al., 2023) are properly dealt with.

5. Conclusion

Argument quality assessment has become a core task in NLP research on computational argumentation, due to its importance for various applications, from debating technologies and argument search to discussion moderation and writing support. However, a reliable assessment is often hampered by the diversity of quality notions involved and the subjectivity of their perception. In this survey-based position paper, we have raised the question of how to drive research on instruction-following large language models (LLMs) for argument quality to substantially evolve the state of the art.

Our survey of 83 recent papers confirms that argument quality research often targets conceptual quality notions and the factors that influence these notions, aside from the computational assessment and improvement of argument quality. We have argued that many limitations of prior work can be overcome, if LLMs are not just simply prompted for argument quality assessment, but if systematic ways to instruct LLMs for argument quality during instruction fine-tuning are found. This is due to the fact that instruction-following LLMs, for the first time in machine learning-based NLP research, make the connection between inputs and outputs of tasks explicit, namely, through the instructions. Thereby, all knowledge that an LLM has processed during pretraining and fine-tuning can be shared across tasks and contexts.

To guide future work in this direction, we have delineated a blueprint of how to approach the instruction fine-tuning process. Realizations of this process will likely bring up further problems, not all are foreseeable at this point. Moreover, LLMs that effectively predict human perception of argument quality directly raise concerns, as detailed in our ethics statement below. Still, we are confident that coordinated efforts towards sustainable research on LLMs for argument quality will enable the community to progress on core visions of computational argumentation—whether it is about ways to overcome filter bubbles or about the individualized support of argumentation learners. The paper at hand seeks to lay the ground for this research.

6. Ethics Statement

Despite the huge potential of instruction-following LLMs for argument quality assessment across various applications of computational argumentation outlined in Section 4, the blueprint from Section 3 also comes with limitations and ethical concerns. We acknowledge and analyze these in this section.

6.1. Limitations

The discussed potential we see is based on our survey of argument quality research (Section 2), initial works of the emerging body of instruction fine-tuning research (e.g., Peng et al., 2023), and our own preliminary tests. Yet, the work at hand remains a position paper, meaning that experimental research still needs to establish whether the outlined blueprint or similar paths will actually result in substantial progress. It is possible that argument-specific instruction fine-tuning of large language models (LLMs) does not improve over the capabilities of a general large-scale tuning. Also, the systematic ways that we have proposed to establish above remain to be found; there is no obvious way of directly obtaining them. This challenge is in line with the overall state of instruction fine-tuning research, both in academia and in industry.

Regarding the specific challenges of argument quality raised in Section 1, another limitation refers to general possibility that information required to achieve a realiable assessment is simply not available, due to specificities of the setting or underlying privacy regulations. This particularly includes the audience whose quality perception is to be represented, but possibly also aspects of the (temporal, geographical, and social) context in which an argument is to be considered. Also, as soon as we rely on human-created training data for instruction fine-tuning, the creators' biases and values affect its impact. Ultimately, we cannot expect LLMs to tackle a task reliably under conditions that simply do not suffice to make an informed judgment.

6.2. Ethical Concerns

Many arising ethical issues of the use of LLMs for argument quality assessment are general and not specific to computational argumentation, such as the increased environmental impact of bigger models, privacy issues, hallucinations, the potential of models to encode unfair exclusive (Dev et al., 2021; Lauscher et al., 2022a) and stereotypical biases (Blodgett et al., 2020), which may result in allocational and representational harms (Barocas et al., 2017). However, we believe that some of them deserve specific attention in scenarios where argument quality is assessed or optimized, particularly when leveraging the power of LLMs.

In particular, argument quality assessment may be used in sensitive applications such as digital education; for example, to support argumentative writing or to provide guidance on political opinion formation. For such applications, factual errors are particularly problematic, as they may easily lead to wrong or shifted beliefs. Whenever LLMs may generate argumentative content, say, for debating technologies or to fill gaps in argument search as discussed, extra measures should thus be taken to prevent hallucinations. We have sketched how to generally account for them in Section 3, but fully avoiding them may be hard given how LLMs work.

Similarly, unfair social biases are easy to perpetuate in such applications, since the output of LLMs for argument quality assessment will often directly affect human views. This raises various integral and partly self-referential questions, such as who decides on what makes a good argument, or, how to decide on the ethical uses of instruction-following LLMs in argument quality assessment? We expect that universally-accepted answers to these questions may not exist, as they also depend on the values within a culture or society.

As the limitations discussed above imply, further ethical concerns refer to the tension between the inclusion of audience and debater information for a more accurate quality assessment. While an argument's persuasive effect is, for instance, highly dependent on the sociodemographic aspects of its audience, it is questionable in general to what extent an application of respective methods should have access to personal data. Such aspects need to be handled with care, and under consultation with an ethics board, where needed.

For a successful and societal beneficial use of instruction-following LLMs, we thus conclude that future research on argument quality assessment needs to find answers to such questions and to proactively raise and discuss them explicitly.

7. Acknowledgments

This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG) within the project OASiS, project number 455913891, as part of the Priority Program “Robust Argumentation Machines (RATIO)” (SPP-1999), as well as within the project ArgSchool, project number 453073654. It was also partially funded by the Bundesministerium für Bildung und Forschung (BMBF) within the project E-DELIB, project number 01IS20050, and under the Excellence Strategy of the German Federal Government and the States. This work has also been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR- 19-P3IA-0002.

8. Bibliographical 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.

Milad Alshomary, Roxanne El Baff, Timon Gurcke, and Henning Wachsmuth. 2022. The moral debater: A study on the computational generation of morally framed arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8782–8797, Dublin, Ireland. Association for Computational Linguistics.

Lisa P. Argyle, Ethan Busby, Joshua Gubler, Chris Bail, Thomas Howe, Christopher Rytting, and David Wingate. 2023. AI chat assistants can improve conversations about divisive topics.

Aristotle. ca. 350 B.C.E./ translated 2007. *On Rhetoric: A Theory of Civic Discourse*. Oxford University Press, Oxford, UK. Translated by George A. Kennedy.

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.

Roy Bar-Haim, Liat Ein-Dor, Matan Orbach, Elad Venezian, and Noam Slonim. 2021. Advances in debating technologies: Building AI that can debate humans. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 1–5, Online. Association for Computational Linguistics.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *Proceedings of 9th Annual Conference of the Special Interest Group for Computing, Information and Society*.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2023. How (not) to use sociodemographic information for subjective NLP tasks. *CoRR*, abs/2309.07034.

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.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Katarzyna Budzynska and Chris Reed. 2019. Advances in argument mining. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 39–42, Florence, Italy. 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. 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.

Tuhin Chakrabarty, Christopher Hidey, and Smaranda Muresan. 2021. ENTRUST: Argument reframing with language models and entailment. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4958–4971, Online. Association for Computational Linguistics.

Lisa A. Chalaguine and Anthony Hunter. 2020. A Persuasive Chatbot Using a Crowd-Sourced Argument Graph and Concerns. In *Computational*

Models of Argument. Proceedings of COMMA 2020, volume 326 of *Frontiers in Artificial Intelligence and Applications*, pages 9–20. IOS Press.

Wei-Fan Chen, Mei-Hua Chen, Garima Mudgal, and Henning Wachsmuth. 2022a. Analyzing culture-specific argument structures in learner essays. In *Proceedings of the 9th Workshop on Argument Mining*, pages 51–61, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Zaiqian Chen, Daniel Verdi do Amarante, Jenna Donaldson, Yohan Jo, and Joonsuk Park. 2022b. 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.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yuning Ding, Marie Bexte, and Andrea Horbach. 2023. Score it all together: A multi-task learning study on automatic scoring of argumentative essays. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13052–13063, Toronto, Canada. Association for Computational Linguistics.

Lorik Dumani, Patrick J. Neumann, and Ralf Schenkel. 2020. A framework for argument retrieval. In *Advances in Information Retrieval*, pages 431–445, Cham. Springer International Publishing.

Esin Durmus and Claire Cardie. 2018. Exploring the role of prior beliefs for argument persuasion. 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 1035–1045, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Ryo Egawa, Gaku Morio, and Katsuhide Fujita. 2019. Annotating and analyzing semantic role of elementary units and relations in online persuasive arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 422–428, Florence, Italy. Association for Computational Linguistics.

Roxanne El Baff, Khalid Al Khatib, Benno Stein, and Henning Wachsmuth. 2020a. Persuasiveness of news editorials depending on ideology and personality. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 29–40, Barcelona, Spain (Online). 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. 2020b. 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, Dannica Fleuß, and Sarah-Michelle Nienhaus. 2021. Different arenas, different deliberative quality? using a systemic framework to evaluate online deliberation on immigration policy in germany. *Policy & Internet*, 13(1):86–112.

Neele Falk, Iman Jundi, Eva Maria Vecchi, and Gabriella Lapesa. 2021. Predicting moderation of deliberative arguments: Is argument quality the key? In *Proceedings of the 8th Workshop on Argument Mining*, pages 133–141, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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. Bridging argument quality and deliberative quality annotations with adapters. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2469–2488, Dubrovnik, Croatia. Association for Computational Linguistics.

Marc Feger, Jan Steimann, and Christian Meter. 2020. Structure or Content? Towards Assessing Argument Relevance. In *Computational Models of Argument. Proceedings of COMMA 2020*, volume 326 of *Frontiers in Artificial Intelligence and Applications*, pages 203–214. IOS Press.

Michael Fromm, Max Berrendorf, Evgeniy Faerman, and Thomas Seidl. 2023. Cross-domain argument quality estimation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13435–13448, Toronto, Canada. Association for Computational Linguistics.

Michael Fromm, Evgeniy Faerman, Max Berrendorf, Siddharth Bhargava, Ruoxia Qi, Yao Zhang, Lukas Dennert, Sophia Selle, Yang Mao, and Thomas Seidl. 2021. Argument mining driven analysis of peer-reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 6, pages 4758–4766. AAAI Press.

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.

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.

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. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 7805–7813. AAAI.

Yunfan Gu, Zhongyu Wei, Maoran Xu, Hao Fu, Yang Liu, and Xuanjing Huang. 2018. Incorporating topic aspects for online comment convincingness evaluation. In *Proceedings of the 5th Workshop on Argument Mining*, pages 97–104, Brussels, Belgium. Association for Computational Linguistics.

Zhen Guo and Munindar P. Singh. 2023. Representing and determining argumentative relevance in online discussions: A general approach. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):292–302.

Timon Gurcke, Milad Alshomary, and Henning Wachsmuth. 2021. Assessing the sufficiency of arguments through conclusion generation. In *Proceedings of the 8th Workshop on Argument Mining*, pages 67–77, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. 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. Association for Computational Linguistics.

Md Kamrul Hasan, James Spann, Masum Hasan, Md Saiful Islam, Kurtis Haut, Rada Mihalcea, and Ehsan Hoque. 2021. Hitting your MARQ: Multimodal ARgument quality assessment in long debate video. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6387–6397, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ethan Haworth, Ted Grover, Justin Langston, Ankush Patel, Joseph West, and Alex C. Williams. 2021. Classifying reasonability in retellings of personal events shared on social media: A preliminary case study with /r/amitheasshole. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):1075–1079.

Philipp Heinisch, Moritz Plenz, Juri Opitz, Anette Frank, and Philipp Cimiano. 2022. Data augmentation for improving the prediction of validity and novelty of argumentative conclusions. In *Proceedings of the 9th Workshop on Argument Mining*, pages 19–33, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Carolin Holtermann, Anne Lauscher, and Simone Ponzetto. 2022. Fair and argumentative language modeling for computational argumentation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7841–7861, Dublin, Ireland. Association for Computational Linguistics.

Kuo-Yu Huang, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. Hargan: Heterogeneous argument attention network for persuasiveness prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):13045–13054.

Mana Ihori, Hiroshi Sato, Tomohiro Tanaka, and Ryo Masumura. 2022. Multi-perspective document revision. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6128–6138, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ayush Jain and Shashank Srivastava. 2021. Does social pressure drive persuasion in online fora? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9201–9208, Online and Punta Cana, Dominican Republic. 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, Shivani Poddar, Byungsoo Jeon, Qinlan Shen, Carolyn Rosé, and Graham Neubig. 2018. Attentive interaction model: Modeling changes in view in 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 103–116, New Orleans, Louisiana. 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.

Zixuan Ke, Winston Carlile, Nishant Gurrapadi, and Vincent Ng. 2018. Learning to give feedback: Modeling attributes affecting argument persuasiveness in student essays. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4130–4136. International Joint Conferences on Artificial Intelligence Organization.

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.

Nataliia Kees, Michael Fromm, Evgeniy Faerman, and Thomas Seidl. 2021. Active learning for argument strength estimation. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 144–150, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Johannes Kiesel, Kevin Lang, Henning Wachsmuth, Eva Hornecker, and Benno Stein. 2020. Investigating expectations for voice-based and conversational argument search on the web. In *Proceedings of the 2020 Conference on Human Information Interaction & Retrieval (CHIIR 2020)*, CHIIR '20, pages 53–62, New York, NY, USA. Association for Computing Machinery.

Mark Klein. 2012. Enabling large-scale deliberation using attention-mediation metrics. *Computer Supported Cooperative Work (CSCW)*, 21:449–473.

Jonathan Kobbe, Ines Rehbein, Ioana Hulpuș, and Heiner Stuckenschmidt. 2020. Exploring morality in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 30–40, Online. Association for Computational Linguistics.

Varada Kolhatkar and Maite Taboada. 2017a. Constructive language in news comments. In *Proceedings of the First Workshop on Abusive Language Online*, pages 11–17, Vancouver, BC, Canada. Association for Computational Linguistics.

Varada Kolhatkar and Maite Taboada. 2017b. Using New York Times picks to identify constructive comments. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 100–105, Copenhagen, Denmark. Association for Computational Linguistics.

Takahiro Kondo, Koki Washio, Katsuhiko Hayashi, and Yusuke Miyao. 2021. Bayesian argumentation-scheme networks: A probabilistic model of argument validity facilitated by argumentation schemes. In *Proceedings of the 8th Workshop on Argument Mining*, pages 112–124, Punta Cana, Dominican Republic. 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.

Gabriella Lapesa, Eva Maria Vecchi, Serena Villata, and Henning Wachsmuth. 2023. Mining, assessing, and improving arguments in NLP and the social sciences. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–6, Dubrovnik, Croatia. Association for Computational Linguistics.

Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022a. Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1221–1232, Gyeongju, Republic of Korea. International Committee on 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. 2022b. 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. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Jialu Li, Esin Durmus, and Claire Cardie. 2020. Exploring the role of argument structure in online debate persuasion. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8905–8912, Online. Association for Computational Linguistics.

Jingjing Li, Zichao Li, Tao Ge, Irwin King, and Michael R. Lyu. 2022. Text revision by on-the-fly representation optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10956–10964.

Xin Liu, Jiefu Ou, Yangqiu Song, and Xin Jiang. 2021. Exploring discourse structures for argument impact classification. 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 3958–3969, Online. 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.

Liane Longpre, Esin Durmus, and Claire Cardie. 2019. Persuasion of the undecided: Language vs. the listener. In *Proceedings of the 6th Workshop on Argument Mining*, pages 167–176, Florence, Italy. Association for 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. Association for Computational Linguistics.

Kelvin Luu, Chenhao Tan, and Noah A. Smith. 2019. Measuring online debaters' persuasive skill from text over time. *Transactions of the Association for Computational Linguistics*, 7:537–550.

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.

Tobias Mayer, Santiago Marro, Elena Cabrio, and Serena Villata. 2021. Enhancing evidence-based medicine with natural language argumentative analysis of clinical trials. *Artif. Intell. Medicine*, 118:102098.

Farjana Sultana Mim, Naoya Inoue, Paul Reisert, Hiroki Ouchi, and Kentaro Inui. 2019. Unsupervised learning of discourse-aware text representation for essay scoring. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 378–385, Florence, Italy. 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.

Huy Nguyen and Diane Litman. 2018. Argument mining for improving the automated scoring of persuasive essays. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

OpenAI. 2023. GPT-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Chan Young Park, Julia Mendelsohn, Karthik Radhakrishnan, Kinjal Jain, Tushar Kanakagiri, David Jurgens, and Yulia Tsvetkov. 2021. Detecting community sensitive norm violations in online conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3386–3397.

Joonsuk Park, Cheryl Blake, and Claire Cardie. 2015. Toward machine-assisted participation in erulemaking: An argumentation model of evaluability. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, ICAIL '15, pages 206–210, New York, NY, USA. ACM.

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).

Joonsuk Park, Sally Klingel, Claire Cardie, Mary Newhart, Cynthia Farina, and Joan-Josep Vallbé. 2012. Facilitative moderation for online participation in eRulemaking. In *Proceedings of the 13th Annual International Conference on Digital Government Research*. ACM.

Amalie Pauli, Leon Derczynski, and Ira Assent. 2022. Modelling persuasion through misuse of rhetorical appeals. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 89–100, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with GPT-4.

Isaac Persing and Vincent Ng. 2017a. Lightly-supervised modeling of argument persuasiveness. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 594–604, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Isaac Persing and Vincent Ng. 2017b. 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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Peter Potash, Robin Bhattacharya, and Anna Rumshisky. 2017. Length, interchangeability, and external knowledge: Observations from predicting argument convincingness. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 342–351, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Martin Potthast, Lukas Gienapp, Florian Euchner, Nick Heilenkötter, Nico Weidmann, Henning Wachsmuth, Benno Stein, and Matthias Hagen. 2019. Argument search: Assessing argument relevance. In *42nd International ACM Conference on Research and Development in Information Retrieval (SIGIR 2019)*, pages 1117–1120. ACM.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212, Online. Association for Computational Linguistics.

Niklas Rach, Yuki Matsuda, Johannes Daxenberger, Stefan Ultes, Keiichi Yasumoto, and Wolfgang Minker. 2020. Evaluation of argument search approaches in the context of argumentative dialogue systems. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 513–522, Marseille, France. European Language Resources Association.

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.

Ekaterina Saveleva, Volha Petukhova, Marius Mosbach, and Dietrich Klakow. 2021. Graph-based argument quality assessment. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1268–1280, Held Online. INCOMA Ltd.

Tsukasa Shiota and Kazutaka Shimada. 2022. Annotation and multi-modal methods for quality assessment of multi-party discussion. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 175–182, Manila, Philippines. Association for Computational Linguistics.

Ruth Shortall, Anatol Itten, Michiel van der Meer, Pradeep Murukannaiah, and Catholijn Jonker. 2022. Reason against the machine? Future directions for mass online deliberation. *Frontiers in Political Science*, 4:946589.

Edwin Simpson and Iryna Gurevych. 2018. Finding convincing arguments using scalable Bayesian preference learning. *Transactions of the Association for Computational Linguistics*, 6:357–371.

Gabriella Skitalinskaya, Jonas Klaff, and Henning Wachsmuth. 2021. Learning from revisions: Quality assessment of claims in argumentation at scale. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1718–1729, Online. Association for Computational Linguistics.

Gabriella Skitalinskaya, Maximilian SpliethÃ¶ver, and Henning Wachsmuth. 2023. Claim optimization in computational argumentation. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 134–152, Prague, Czechia. Association for Computational Linguistics.

Gabriella Skitalinskaya and Henning Wachsmuth. 2023. To revise or not to revise: Learning to detect improvable claims for argumentative writing support. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15799–15816, Toronto, Canada. Association for Computational Linguistics.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, et al. 2021. An autonomous debating system. *Nature*, 591(7850):379–384.

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 2020*, pages 3875–3881. ijcai.org.

Maximilian Spliethöver and Henning Wachsmuth. 2020. Argument from old man's view: Assessing social bias in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 76–87, Online. Association for Computational Linguistics.

Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018. ArgumenText: Searching for arguments in heterogeneous sources. In *Proceedings of the 2018 NAACL: Demonstrations*, pages 21–25.

Manfred Stede and Jodi Schneider. 2018. *Argumentation Mining*. Number 40 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool.

Shahbaz Syed, Roxanne El Baff, Johannes Kiesel, Khalid Al Khatib, Benno Stein, and Martin Potthast. 2020. News editorials: Towards summarizing long argumentative texts. In *Proceedings of* *the 28th International Conference on Computational Linguistics*, pages 5384–5396, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Shahbaz Syed, Timon Ziegenbein, Philipp Heinisch, Henning Wachsmuth, and Martin Potthast. 2023. Frame-oriented summarization of argumentative discussions. In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, pages 114–129, Prague, Czechia. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

Terne Sasha Thorn Jakobsen, Maria Barrett, and Anders Søgaard. 2021. Spurious correlations in cross-topic argument mining. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 263–277, Online. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Michiel van der Meer, Myrthe Reuver, Urja Khurana, Lea Krause, and Selene Baez Santamaria. 2022. Will it blend? mixing training paradigms & prompting for argument quality prediction. In *Proceedings of the 9th Workshop on Argument Mining*, pages 95–103, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Frans H. van Eemeren and Rob Grootendorst. 2004. *A Systematic Theory of Argumentation: The Pragma-Dialectical Approach*. Cambridge University Press, Cambridge, UK.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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, 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. Association for Computational Linguistics.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Alberdingk Tim 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. Association for Computational Linguistics.

Henning Wachsmuth, Martin Potthast, Khalid Al-Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017c. Building an argument search engine for the web. In *Proceedings of the 4th Workshop on Argument Mining*, pages 49–59. 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.

Lu Wang, Nick Beauchamp, Sarah Shugars, and Kechen Qin. 2017. Winning on the merits: The joint effects of content and style on debate outcomes. *Transactions of the Association for Computational Linguistics*, 5:219–232.

Xiaoou Wang, Elena Cabrio, and Serena Villata. 2023a. Argument and counter-argument generation: A critical survey. In *Natural Language*

Processing and Information Systems, pages 500–510, Cham. Springer Nature Switzerland.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-instruct: Aligning language models with self-generated instructions.

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.

Wonsuk Yang, Seungwon Yoon, Ada Carpenter, and Jong Park. 2019. 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. Association for Computational Linguistics.

Xinran Zhao, Esin Durmus, Hongming Zhang, and Claire Cardie. 2021. Leveraging topic relatedness for argument persuasion. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4401–4407, Online. 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, Toronto, Canada. Association for Computational Linguistics.

A. List of Surveyed Papers

Table A shows the full list of all 83 surveyed papers that resulted from our search and filtering process in Section 2, along with the primary research direction and main aspect of each paper and the research community where the respective paper has been published: natural language processsing (NLP), artificial intelligence (AI), information retrieval (IR), or computational argumentation (CA). The research directions and aspects are detailed in Section 2 as well as selected papers from the list.

#	Research Direction	Main Aspect	Paper	Community
1	**Conceptual Notions**	Notions of Maximal Quality	Atkinson et al. (2019)	NLP
2			Beigman Klebanov et al. (2017)	NLP
3			Dumani et al. (2020)	IR
4			El Baff et al. (2018)	NLP
5			Falk et al. (2021)	NLP
6			Gretz et al. (2020)	AI
7			Guo and Singh (2023)	AI
8			Joshi et al. (2023)	NLP
9			Ke et al. (2019)	NLP
10			Kolhatkar and Taboada (2017b)	NLP
11			Lauscher et al. (2020)	NLP
12			Ng et al. (2020)	NLP
13			Nguyen and Litman (2018)	AI
14			Potthast et al. (2019)	IR
15			Shiota and Shimada (2022)	NLP
16			Wachsmuth et al. (2017a)	NLP
17		Notions of Minimal Quality	Habernal et al. (2018)	NLP
18			Haworth et al. (2021)	AI
19			Kolhatkar and Taboada (2017a)	NLP
20			Jin et al. (2022)	NLP
21			Park and Cardie (2018)	NLP
22			Pauli et al. (2022)	NLP
23			Persing and Ng (2017b)	AI
24			Ziegenbein et al. (2023)	NLP
25	**Influence Factors**	Argument-related Factors	Carlile et al. (2018)	NLP
26			Chen et al. (2022a)	NLP
27			Durmus et al. (2019)	NLP
28			Egawa et al. (2019)	NLP
29			El Baff et al. (2020b)	NLP
30			Kobbe et al. (2020)	NLP
31			Li et al. (2020)	NLP
32			Luu et al. (2019)	NLP
33			Persing and Ng (2017a)	NLP
34			Potash et al. (2017)	NLP
35			Skitalinskaya et al. (2021)	NLP
36			Toledo et al. (2019)	NLP
37			Wachsmuth and Werner (2020)	NLP
38			Wang et al. (2017)	NLP
39			Wambsganss and Niklaus (2022)	NLP
40			Zhao et al. (2021)	NLP
41		Context-related Factors	Al Khatib et al. (2020)	NLP
42			Alshomary et al. (2022)	NLP
43			Durmus and Cardie (2018)	NLP
44			Durmus and Cardie (2019)	NLP
45			El Baff et al. (2020a)	NLP
46			Fromm et al. (2023)	NLP
47			Gu et al. (2018)	NLP
48			Hasan et al. (2021)	NLP
49			Kornilova et al. (2022)	NLP
50			Jain and Srivastava (2021)	NLP
51			Liu et al. (2022)	NLP
52			Longpre et al. (2019)	NLP
53			Lukin et al. (2017)	NLP
54			Wiegmann et al. (2022)	NLP
55	**Computational Models**	Models for Assessment	Ding et al. (2023)	NLP
56			Falk and Lapesa (2022)	NLP
57			Falk and Lapesa (2023)	NLP
58			Feger et al. (2020)	CA
59			Gienapp et al. (2020)	NLP
69			Gleize et al. (2019)	NLP
61			Gurcke et al. (2021)	NLP
62			Holtermann et al. (2022)	NLP
63			Huang et al. (2021)	AI
64			Jo et al. (2018)	NLP
65			Kondo et al. (2021)	NLP
66			Liu et al. (2021)	NLP
67			Marro et al. (2022)	NLP
68			Mim et al. (2019)	NLP
69			Saveleva et al. (2021)	NLP
70			Simpson and Gurevych (2018)	NLP
71			Song et al. (2020)	AI
72			van der Meer et al. (2022)	NLP
73		Models for Improvement	Ke et al. (2018)	AI
74			Skitalinskaya and Wachsmuth (2023)	NLP
75			Skitalinskaya et al. (2023)	NLP
76	**Other**	Other	Chalaguine and Hunter (2020)	CA
77			Chen et al. (2022b)	NLP
78			Fromm et al. (2021)	AI
79			Heinisch et al. (2022)	NLP
80			Kees et al. (2021)	NLP
81			Kiesel et al. (2020)	IR
82			Rach et al. (2020)	NLP
83			Yang et al. (2019)	NLP

Table 1: The list of all 83 surveyed NLP, AI, IR, and CA papers, ordered by their primary general research direction and main aspect and then by author names and year. 23 papers deal with *conceptual notions* primarily, 31 with *influence factors*, 21 with *computational models*, and eight have *other* primary directions.