Bridging Argument Quality and Deliberative Quality Annotations with Adapters

Neele Falk and **Gabriella Lapesa**
Institute for Natural Language Processing, University of Stuttgart
{neele.falk,gabriella.lapesa}@ims.uni-stuttgart.de

Abstract

Assessing the quality of an argument is a complex, highly subjective task, influenced by heterogeneous factors (e.g., prior beliefs of the annotators, topic, domain, and application), and crucial for its impact in downstream tasks (e.g., argument retrieval or generation). Both the Argument Mining and the Social Science community have devoted plenty of attention to it, resulting in a wide variety of argument quality dimensions and a large number of annotated resources. This work aims at a better understanding of how the different aspects of argument quality relate to each other from a practical point of view. We employ adapterfusion (Pfeiffer et al., 2021) as a multi-task learning framework which a) can improve the prediction of individual quality dimensions by injecting knowledge about related dimensions b) is efficient and modular and c) can serve as an analysis tool to investigate relations between different dimensions. We conduct experiments on 6 datasets and 20 quality dimensions. We find that the majority of the dimensions can be learned as a weighted combination of other quality aspects, and that for 8 dimensions adapter fusion improves quality prediction. Last, we show the benefits of this approach by improving the performance in an extrinsic, out-of-domain task: prediction of moderator interventions in a deliberative forum.

1 Introduction

Although people have been dealing with the art of persuasion since ancient times, there are many answers to the question of what constitutes a good argument or good argumentation, and none can be considered the best: the quality of arguments is complex, subjective and depends on the context in which the quality is assessed and on the prerequisites (attitudes and values) of the one who judges it. Despite the high complexity of this task, a considerable research effort has been done to automatically model argument quality in different contexts (Wachsmuth et al., 2017a) due to its usefulness in downstream applications, such as automatic writing assistants (Wambsganss et al., 2020), argument extraction (Alshomary et al., 2021) and generation (Gurcke et al., 2021). Social Science offers another whole field of theories and definitions about argument quality in which the focus is usually not only on the argument itself but on the discussion between participants thus emphasizing the deliberative goal of the discourse (Gerber et al., 2018).

These two research communities, Argument Mining (AM) and Deliberative Theory (DT), have not only produced different theories of argument quality, but also a number of annotated datasets on the basis of which the models for the automatic assessment can be (or have been) trained. In both AM and DT, argument quality (AQ) and its Social Science counterpart, deliberative quality (DQ) are broken down into finer-grained dimensions. Such dimensions map, for example, whether an argument is logically constructed (micro-level), or constructive in the context of an overall discussion (macro-level). However, neither individual quality dimensions, nor an aggregated score can do justice to the complexity of this concept. Besides, a model that represents different aspects but has only been trained on one dataset will reproduce data-specific biases and may be less robust on other domains.

Multi-task learning (in this work, each quality dimension, e.g., logical cogency, clarity, persuasiveness, is a task), and the training data drawn from different datasets are a solution, as they allow to integrate the different dimensions and data sources from the two research communities. We propose to implement it using adapters (Houlsby et al., 2019), modules added between the layers of a transformer model. Differently from fine-tuning of the full model, adapters allow to use a minimal amount of parameters while still achieving good performance. Differently from standard multi-task learning, adapters do not require all the tasks to be

Findings of the Association for Computational Linguistics: EACL 2023, pages 2469–2488

learnt simultaneously, but they can be learnt as specific modules that can also be combined (fused or stacked). The modular design of the adapters then allows for flexible composition of the individual quality aspects and thus can be used in various configurations; this property facilitates future research of argument quality in different domains and lends itself as tool for the investigation of the relationship between different quality dimensions, both within and across disciplines. We experiment with 6 datasets containing AQ and DQ annotations, for a total of 20 dimensions (AQ:8; DQ: 12) covering a wide range of logical, rhetorical and dialectical aspects and a variety of domains and topics. Our work proceeds in two steps.

In the first step, we employ adapter fusion to learn a target dimension as a weighted combination of single-task adapters (e.g., clarity as a combination of cogency and effectiveness). We improve the results with respect to single-task learning on 8 dimensions out of 20 in a low-resource scenario. Furthermore, fusion activation patterns provide us a tool to investigate the relationship between different quality dimensions.

In the second step, we employ quality adapters for a new task on a new dataset: predicting that a post in a deliberative forum needs moderation (Park et al., 2012). A fusion based on quality adapters is compared to baseline and full fine-tuning, outperforming both. Moreover, our analysis shows that in solving the task, the models exploit information from all major quality sub-categories. Crucially for the downstream application, casting moderation prediction as a fusion of quality adapters allows us to provide recommendation explained along specific quality dimensions (e.g., "this comment has major issues with the logical side of argumentation and it is disrespectful").

The contributions of this paper are at multiple level: a) at the level of task and methods, this is the first work which employs adapters for finer-grained AQ dimensions; b) at the conceptual/theoretical level, we make a first step in the integration of theories of AQ and DQ, bridging between the annotations produced by the two communities and proposing adapter activation as a tool to empirically compare the conceptual core shared by AQ and DQ dimensions c) at the level of application, we show that quality adapters can support the task of predicting moderation of user comments, additionally contributing a theory-based explanation layer.

2 Related Work

Argument Quality Much work on automatic modeling and annotation of argument quality (AQ) in the Argument Mining community focuses either on a specific aspect of quality (e.g. argument relevance (Wachsmuth et al., 2017b), sufficiency (Stab and Gurevych, 2017)) or a more general notion of argument quality based on human intuition (Habernal and Gurevych, 2016). Wachsmuth et al. (2017a) proposes a holistic taxonomy based on different theories of argument quality, inspired from rhetoric and linguistics, which divides AQ into three main sub-categories. The logical dimension measures whether an argument has premises and a valid conclusion (cogency) thus takes the content and structure of a single argument into account. The rhetorical dimension (effectiveness) measures the persuasiveness of the argument and takes into account *how* it is presented (style, emotional appeal). The dialectical dimension (reasonableness) plays a more important role in the context of a discourse and reflects whether an argument is valid towards a universal audience (e.g. whether the reasoning is based on values generally accepted by the society) or whether it is constructive in helping to resolve issues. Wachsmuth et al. (2017a) construct a corpus consisting of 302 arguments annotated with the three core and 15 sub-dimensions. Wachsmuth and Werner (2020) investigate which linguistic features are predictive of the fine-grained dimensions and which of the dimensions can be automatically assessed based on the textual input representations alone. The work by Fromm et al. (2022) are the first that try to combine AQ definitions from different corpora and based on different annotation schemas into one model. They investigate the generalizability of AQ when combining different sources and explore multi-task learning for assessing it in four different datasets. On top of that they investigate the relationship between AQ and other AM tasks such as evidence detection.

While most work on AQ in the Argument Mining community focuses on the logical dimension or specific aspects of persuasion, research on deliberative quality (DQ) from Social Science puts the discourse as a whole and the interaction between discourse participants into the focus. Here, argument quality (or discourse / deliberative quality) is investigated to find out which tools and solutions (e.g. moderation, platform design, structured overviews) can contribute to a more productive and

respectful public discourse. Thus, the annotated datasets from this domain complement the ones from the AM community providing many aspects of the rhetorical and dialectic dimensions.

Adapters Adapters (Houlsby et al., 2019) are a set of task-specific parameters that are introduced in every layer of a transformer (Vaswani et al., 2017) and updated for a specific task while the rest of the pre-trained language-model parameters is kept frozen. Besides being more efficient than full fine-tuning, adapters can be used as building blocks for other tasks due to their modular architecture and are therefore particularly well suited for transfer– and multi-task learning (He et al., 2021) and to inject external knowledge sources to solve downstream tasks (Lauscher et al., 2020a). Pfeiffer et al. (2021) propose to train task-specific adapters first (knowledge extraction) and combine them in a second step (knowledge composition) using self-attention to mitigate catastrophic interference, a problem which often occurs with traditional multi-task learning approaches. In their work, this approach has proven to be useful especially in low-resource settings which is often the case for complex annotations such as the AQ ones. To the best of our knowledge, our work is the first to employ adapters for AQ to conduct a systematic comparison of AQ and DQ on different data sources.

3 Datasets

For our experiments we rely on diversity, both in terms of data sets and different conceptualizations of argument quality. Therefore, we also integrate two datasets from the Social Sciences, which are not established in the argument quality community, but show a particularly large variety of dimensions.
Europolis (Gerber et al., 2018): consists of transcriptions of a face-to-face discussion about the topic immigration, initiated by the European Union in order to enable deliberation on a European level. The spoken multi-lingual contributions have been transcribed, partially translated and annotated with five different dimensions of DQ by political scientists, each dimension between two to five labels that can be arranged on a scale from a low to a high standard of deliberative abilities. The dimensions capture the logical aspect (*justification*), rhetorical aspects (*storytelling*) and dialectic aspects (*common Good*, *interactivity* and *respect*).
THF/BK (Esau, 2022): this dataset contains comments from two online citizen dialogues on municipal issues: one on the further development of the "Tempelhofer Feld" site in Berlin and the other on the use of the former lignite area in North Rhine-Westphalia. The data was annotated by political scientists with different dimensions of DQ using a binary label for each dimension. The goal of the work was to investigate the relationship between "classic standards of deliberation", such as rationality and constructiveness and alternative forms of deliberation, such as humor, narratives and the use of emotions. This dataset therefore offers annotations for the so far rather underexplored and more affective dimensions of argument quality, such as *positive emotions*, *narration* and *empathy*.
Kialo (Durmus et al., 2019): This dataset was created based on the online discussion platform Kialo `https://www.kialo.com` on which users engage in structured discussions about a certain statement. Users are able to rate the *impact* of an argument given its context. The dataset contains arguments about a large number of different topics together with their impact – a label which aggregates impact votes by all users. Durmus et al. (2019) and Li et al. (2020) report F-macro scores between 0.56 and 0.58 using different transformer-based models.
Grammarly Argument Quality Corpus (GAQ) (Ng et al., 2020): this dataset contains online contributions from four different domains annotated with the coarse-grained levels of the taxonomy introduced by Wachsmuth et al. (2017a) on a five-point scale. Lauscher et al. (2020b) evaluate different systems for automatic prediction of the quality scores, also experimenting with different multi-task architectures showing that multi-task learning can lead to improvements for all dimensions.
IBM-Rank-30k (Gretz et al., 2020): the largest available corpus with AQ annotations has been created based on a large quantity of binary annotations for human-generated arguments. The authors evaluate different methods of aggregating the annotations into a continuous score and conduct experiments on the automatic prediction of these scores with a Pearson correlation of around 0.48 on a test set with unseen topics. Lauscher et al. (2020b) found positive correlations between this aggregated AQ score and automatically generated scores for *cogency*, *effectiveness* and *reasonableness* on this corpus.
SwanRank (Swanson et al., 2015): as one of the first datasets with AQ annotations in the AM community this corpus contains arguments from on-

Dataset	size	genre	topics	mean length
SwanRank	5k	online discussion	gay marriage, gun control, death penalty, evolution	19
GAQ	5k	Debates, CQA, Reviews	diverse	109
IBM-Rank-30k	30k	crowd-sourced arguments	71 common controversial topics	18
Kialo	7k	argument maps	741 topics	23
Europolis	1k	face-to-face deliberation	immigration in Europe	131
THF/BK	1k	online deliberation	Redevelopment Tempelhofer Feld (THF) and lignite mining (BK)	124

Table 1: Overview of the datasets: original size, genre, topics and mean length in tokens of contributions

.

dimension	short description	measured	corpus
overall	general argument quality	score (1-5)	GAQ
cogency	acceptable and sufficient premises to draw a conclusion	score (1-5)	GAQ
reasonableness	contribution to resolution of issues, argument is accepted by universal audience	score (1-5)	GAQ
effectiveness	persuasion, rethorical, emotional appeal	score (1-5)	GAQ
quality	general argument quality	score (0-1)	IBM-Rank-30k
clarity	is it hard or easy to interpret the argument?	score (0-1)	Swanson
justification	rationality, providing reasons, reflection	multi-class (4)	Europolis
respect	empathy or respect towards groups (e.g. immigrants)	multi-class (3)	Europolis
storytelling	personal experience, subjective description of an event or situation	binary	Europolis
interactivity	respect towards other participants, reference to other participants arguments	multi-class (4)	Europolis
common good	taking interests of the broader community or utilitarianism based values (justice, equality) into account	multi-class (3)	Europolis
posEmotion	positive emotions are contained in the utterance	binary	THF/BK
proposal	a statement about what or how something is to be done	binary	THF/BK
narration	personal experience, subjective description of an event or situation	binary	THF/BK
reference	participant refers to another discourse participant	binary	THF/BK
argument	providing reasons and/or evidence in favor of or against a claim	binary	THF/BK
negEmotion	negative emotions are contained in the utterance	binary	THF/BK
empathy	Speaker puts themself in the perspective or emotional state of others	binary	THF/BK
Q(uestion) for justification	asks for the reasons for a statement or action	binary	THF/BK
impact	user likes / recommendations	multi-class (3)	Kialo

Table 2: Overview of the datasets with their respective argument quality dimensions

line discussion fora about four controversial topics. The corpus was annotated using crowd-sourcing on a continuous scale expressing whether an argument is easy or hard to interpret, thus reflecting the *clarity* of an argument. More recent experiments on this dataset are for example reported in Gretz et al. (2020) who experiment with fine-tuning transformer-based models after pre-training them on the IBM-Rank-30k dataset.

Table 2 shows an overview of the mentioned datasets and their corresponding quality dimensions, an example with the annotated label / score for each dimension can be found in Tables 13 to 16 in the appendix. Table 1 shows an overview of the six datasets with their respective size and number of topics. While the two datasets from Social Science offer the largest amount of different annotations they are also the smallest in size. On the other hand they consist of full discussions whereas the datasets from Argument Mining consist of single arguments without their broader context.

4 Experiment 1: Modeling AQ and DQ using adapters and adapter-fusion

In the following experiment we are interested in the relationship between different conceptualizations of AQ and DQ from a modeling perspective: does injecting knowledge about other dimensions help to improve the predictions on a target dimension? If so, which dimensions are especially helpful? To investigate this we treat each of the 20 dimensions as a task which we aim to model. We want to compare how the models perform without external information (using only a single-task adapter) with those using information about other dimensions (using multi-task learning with adapter-fusion).

4.1 Experimental setup

The input for all adapter models is the argumentative text, which consists of a sentence, a comment, or a spoken contribution, depending on the data set. We use RoBERTa (Liu et al., 2019) (`roberta-base`) as the backbone transformer model for all dimensions. Note that for each of the 20 single-task adapters we train a task-specific prediction head, depending on the underlying classifi-

cation problem (binary-, multi-class classification or regression). We pick the model with the best results on the validation set (lowest mean-squared error for the regression–, highest F1 macro for the classification tasks using class weights to counteract class imbalance).

Heuristic: how to select source tasks for adapter fusion? As the number of existing dimensions is large (20) we apply a heuristic to select different pools of source tasks for a target quality dimension. For this, we use predictions of dimension-specific adapters as proxies to uncover relationships between different quality aspects. We train an adapter for each dimension on the whole corresponding source dataset, generate predictions on all other datasets and measure pair-wise correlations across the datasets. We hypothesize that dimensions that have a clear positive or negative correlation to the target dimension will be most useful to support modeling that quality aspect, thus we add a dimension as source task if the absolute value of the correlation to the target dimension exceeds a threshold.

We sample source tasks from correlations between all 20 dimensions (*fusion corr ALL*), from correlations between dimensions from datasets with a focus on deliberation (Europolis, THF/BK, Kialo: *fusion corr DQ*) and from those originating from established datasets from Argument Mining (GAQ, IBM, SwanRank: *fusion corr AQ*). We use the third quartile of the correlations of the respective dimensions as the threshold in each case (more details and correlation matrix in Appendix Section C and Figure 6). Appendix table 12 displays the output of the selection based on this heuristic. For most of the target dimensions, it indicates a fusion with between 2 and 9 source dimensions: more logical or general dimensions are more often selected (e.g. most frequent source dimension is *justification*, which gets selected for 14 target dimensions). A qualitative inspection of the suggested combinations shows that the heuristic is picking up sensible conceptual patterns. For example, for the target dimension *empathy* the candidates for fusion in the *fusion corr ALL* setup are *negEmotion*, *story* and *narration*.

For each setup we experiment whether we need to add the adapter of the target dimension as source task (*w own adapter*) as it has been done in Pfeiffer et al. (2021) or whether we can learn a target dimension as a weighted combination from other source dimensions (*w/o own adapt.*) As the multi-tasking approach should be most helpful for low-resource scenarios, we down-sample the larger datasets (Kialo, IBM-Rank-30k, GAQ, SwanRank) to 1000 instances. We use the original train/val/test split for IBM-Rank-30k, GAQ and Kialo and create our own split for THF/BK, SwanRank and Europolis. We train the fusions similar to the single-task adapters with a lower number of epochs and a smaller learning rate ($5e-5$).[1] We train each model with 3 different seeds and report mean and standard deviation of F1 macro score and Pearson correlation in Table 11.

5 Results

Can we improve modeling AQ with adapter-fusion? Table 3 shows the results comparing single-task adapters with the fusion-based models, averaged over three seeds. We use the Almost Stochastic Order test (Del Barrio et al., 2018; Dror et al., 2019) as implemented by Ulmer et al. (2022) to identify for which dimensions multi-task learning can lead to significant improvements.[2]

Our results show that: a) Information about related quality dimensions can improve modeling for individual dimensions (significant improvements for 8 of 20 dimensions). These stem from 4 different datasets, so the trend holds across different datasets from both communities (AM and DT). b) For most dimensions the fusion does not lead to performance drops, which confirms the fact that adapter-fusion is more robust than traditional multi-task learning (no catastrophic forgetting / interference). The individual modules for different dimensions can thus be tried out without major disadvantages for new data sets or quality annotations. c) we gain improvements, even when the target adapter is not provided to the fusion (GAQ dimensions, *narration* and *argumentative*). Thus the target dimension can be learned as a weighted combination of source dimensions that are different. This can be especially useful when we only have little or noisy data for the target dimension available.

[1] For implementation details refer to Appendix Section A.

[2] The test compares two score distributions by quantifying to which extend stochastic order is being violated. If the amount of violation is small enough, one model can be considered as superior (stochastically dominant) over the other.

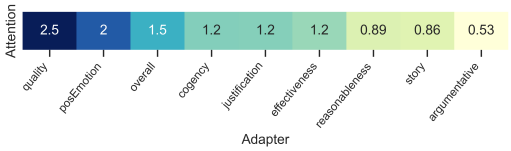
dimension	ST	fusion corr ALL		fusion corr DQ		fusion corr AQ	
		w own adapt.	w/o own adapt.	w own adapt.	w/o own adapt.	w own adapt.	w/o own adapt.
overall	$0.63^{\pm 0.01}$	$0.64^{\pm 0.02}$	$0.64^{\pm 0.02}$			$0.61^{\pm 0.06}$	$\mathbf{0.65^{\pm 0.02}}*$
cogency	$0.41^{\pm 0.10}$	$\mathbf{0.47^{\pm 0.02}}*$	$0.45^{\pm 0.05}$			$\mathbf{0.48^{\pm 0.01}}*$	$\mathbf{0.49^{\pm 0.01}}*$
reasonableness	$0.56^{\pm 0.03}$	$0.55^{\pm 0.03}$	$0.55^{\pm 0.05}$			$0.57^{\pm 0.02}$	$0.56^{\pm 0.04}$
effectiveness	$0.49^{\pm 0.13}$	$\mathbf{0.59^{\pm 0.02}}**$	$\mathbf{0.57^{\pm 0.02}}*$			$\mathbf{0.57^{\pm 0.02}}*$	$\mathbf{0.58^{\pm 0.01}}**$
quality	$0.38^{\pm 0.16}$	$0.48^{\pm 0.05}$	$0.43^{\pm 0.04}$			$0.45^{\pm 0.06}$	$0.43^{\pm 0.07}$
clarity	$0.64^{\pm 0.01}$	$0.63^{\pm 0.03}$	$0.63^{\pm 0.01}$				
justification	$0.46^{\pm 0.04}$	$0.45^{\pm 0.03}$	$0.45^{\pm 0.02}$	$0.46^{\pm 0.03}$	$0.46^{\pm 0.02}$		
story	$0.75^{\pm 0.02}$	$0.76^{\pm 0.02}$	$0.74^{\pm 0.04}$	$0.75^{\pm 0.03}$	$0.73^{\pm 0.04}$		
interactivity	$0.35^{\pm 0.05}$			$\mathbf{0.39^{\pm 0.02}}*$	$0.36^{\pm 0.04}$		
cgood	$0.60^{\pm 0.04}$			$0.61^{\pm 0.05}$	$0.60^{\pm 0.02}$		
posEmotion	$0.64^{\pm 0.03}$	$0.63^{\pm 0.03}$	$0.61^{\pm 0.03}$	$0.64^{\pm 0.03}$	$0.60^{\pm 0.01}$		
proposal	$0.79^{\pm 0.01}$	$0.80^{\pm 0.03}$	$0.79^{\pm 0.02}$	$0.79^{\pm 0.02}$	$0.78^{\pm 0.02}$		
narration	$0.76^{\pm 0.02}$	$0.76^{\pm 0.01}$	$0.77^{\pm 0.01}$	$0.77^{\pm 0.02}$	$\mathbf{0.78^{\pm 0.02}}*$		
reference	$0.80^{\pm 0.01}$	$0.80^{\pm 0.02}$	$0.80^{\pm 0.02}$	$0.81^{\pm 0.01}$	$0.80^{\pm 0.01}$		
argumentative	$0.77^{\pm 0.01}$	$0.77^{\pm 0.02}$	$\mathbf{0.78^{\pm 0.02}}*$	$0.76^{\pm 0.03}$	$0.76^{\pm 0.01}$		
negEmotion	$0.70^{\pm 0.01}$	$\mathbf{0.72^{\pm 0.04}}*$	$0.70^{\pm 0.02}$	$\mathbf{0.71^{\pm 0.01}}*$	$0.70^{\pm 0.02}$		
empathy	$0.69^{\pm 0.04}$	$0.71^{\pm 0.02}$	$0.69^{\pm 0.04}$	$0.69^{\pm 0.03}$	$0.67^{\pm 0.02}$		
Qjustification	$0.89^{\pm 0.01}$			$0.89^{\pm 0.01}$	$0.87^{\pm 0.01}$		
impact	$0.47^{\pm 0.02}$	$\mathbf{0.49^{\pm 0.02}}*$	$0.47^{\pm 0.01}$				

Table 3: Comparison between task-specific adapter and fusions. Average performance (F1 macro and pearson correlation) on the test set. * denotes almost stochastic dominance ($\epsilon_{\text{min}} < \tau$ with $\tau = 0.5$) and ** denotes truly stochastic dominance ($\epsilon_{\text{min}} < \tau$ with $\tau = 0.0$)

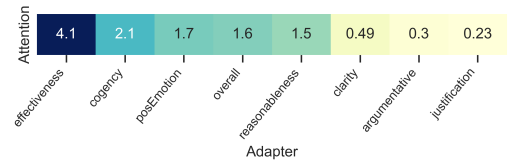
6 Analysis: relationship between AQ and DQ dimensions

For each target dimension we analyze which adapters get activated during inference. We extract the attention scores for source dimensions for each target dimension based on the test set. Similar to Pfeiffer et al. (2021) we assume that high activations indicate more useful source tasks.

General AQ / AQ based on intuition First we compare two very general conceptualizations of general AQ: *quality* (from the IBM-rank dataset) which was trained on a wide variety of controversial topics and *clarity* with a slightly more tolerant conceptualization of quality (is the argument clear / understandable?) on 4 different topics. Both conceptualizations are rather under-specified and based on human intuition, we can thus gain insights into which dimensions play a particularly important role for the intuitive understanding of AQ. Figure 1 visualizes the most activated dimensions. For both dimensions different aspects, logical (*justification*, *cogency*) and rhetorical (*effectiveness*) are activated. Emotions play a role (high activation for *posEmotion*) and all dimensions from GAQ receive high activation indicating that they provide useful information in general. Interestingly, the adapter for *quality* (IBM) gets the most activation when modeling *clarity*, while the other way around is not the case. This may indicate that *clarity* represents a somewhat more specific conceptualization of argument quality, while quality reflects a more general.



(a) clarity (SwanRank)



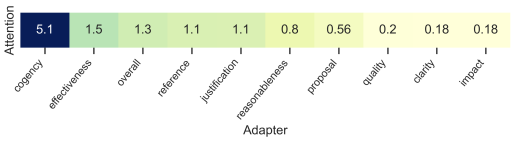
(b) quality (IBM)

Figure 1: General conceptualizations of quality: sum of adapter activations over all layers.

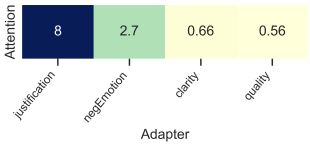
Recall also that the source corpus for *quality* is IBM rank, which covers 71 topics thus resulting in representations that are more applicable to corpora of other domains. Dialectical dimensions are less relevant, as both datasets contain single arguments without a discussion context.

Logical aspect of AQ and DQ With Figure 2 we can compare two logical conceptualizations, one from the AM community (*cogency* from GAQ) and one from the Social Science (*argumentative* from THF/BK). This allows us to explore the extent to which a similar conceptualization of logical argument quality varies between the two datasets from the different research communities.

Argumentative benefits mostly form the other logical dimension of the DQ dataset (*justification*),



(a) cogency (GAQ)



(b) argumentative (THF)

Figure 2: Logical conceptualizations of quality: sum of adapter activations over all layers.

while *cogency* benefits mostly from other dimensions from the same dataset. However *justification* also provides useful information for *cogency* hence seems to be the connecting element between the two conceptualizations. Other useful dimension from a deliberative source are *references* to other people for *cogency* and *negative emotions* for *argumentative*. Having a look at concrete correlation values reveals that the models pick up on positive and negative correlations: arguments with high cogency are less likely to focus on interaction with other people (refer to other peoples arguments) while more argumentative arguments in the THF/BK corpus express more negative sentiment.

Rhetorical aspects: narratives as alternative form of deliberation Finally, we examine a deliberative dimension that is rather rarely studied in the context of argument quality: *narration* (Figure 3). Moreover, this represents a rhetorical quality dimension, which enables us to compare how this kind of quality dimension differs from logical and general argument quality. Emotions as well as classical argumentative properties play a major role (high activation for positive, negative *Emotions* and *argumentative*), indicating that narration and argumentation are often intertwined. The high activation for *empathy* and *reference* (reference to others) illustrates perspective taking, which is characteristic for narrative. Overall, rhetorical and dialectical aspects play more of a role for this dimension.

We can summarize the following trends: either dimensions that come from similar or the same datasets or conceptually related dimensions are particularly activated. However, we also find empirical evidence that emotions play a role in modeling all

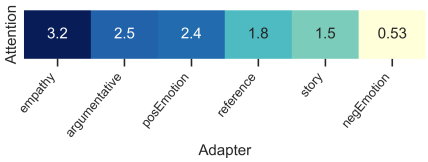


Figure 3: rhetorical conceptualization of quality – narration (THF): sum of adapter activations over all layers.

kinds of dimensions. We suspect that the relationship between emotions and AQ strongly depends on discourse/context, but further research is needed to investigate the relationships more precisely.

7 Experiment 2: predicting moderator interventions

In this experiment, we evaluate the models using a new down-stream application: we want to predict whether a comment in an online discussion should be moderated. Our hypothesis is that we can use information about different quality dimensions to solve the task. Moderation, especially in deliberative discussions such as on civic participation platforms, is a complex task that generally consists of facilitating a productive and fair discussion with respectful interaction. Since the task becomes more difficult for human moderators to perform as the number of participants and comments increases, automatic models can be useful for predicting whether a comment should be moderated. Here, the logical quality dimensions can help distinguish less argumentative from argumentative comments, the rhetorical dimensions are important for ensuring civil interaction, and the dialectical dimensions can identify valuable comments (is a solution proposed or the common good considered?).

We use the dataset from Park et al. (2012), in which the authors annotated the functions of moderation in discussions on a deliberative platform and identified 'quality of comments' as a common reason for intervention. The dataset was used in Falk et al. (2021), who obtained an F1 score of 0.34 using a full fine-tuning approach with roberta-base. The dataset is small and consists of 876 negative and 222 positive instances, a further motivation for a multi-task based approach. We train and test the models on the 5-fold split provided by Falk et al. (2021). As moderator interventions are the minority class we use class weights for all models. We compare the following models:

(quality) scores ST: we generate predictions for

each quality dimension and convert them into scores.[3]. Classifier: logistic regression.
(quality) scores-MT: similar to *quality scores* but we generated with the fusion-based adapters. Classifier: logistic regression.

model	F1 intervention	F1 macro
random baseline	0.29±0.06	0.45 ±0.04
scores-ST	0.37±0.05	0.55±0.04
scores-MT	**0.38±0.04**	0.54±0.04
moderation-ST	0.34±0.03	0.57±0.03
fusion-AQ	0.35±0.04	0.56±0.01
fusion-all	**0.38±0.05**	0.57±0.03
(Falk et al., 2021)	0.34±0.05	0.57±0.03

Table 4: F1 positive (moderator intervention) and F1 macro: average and standard deviation over 5 test sets.

moderation-ST: we train a single-task adapter on the task of moderation intervention.
fusion-AQ: we train a fusion on the task of moderation intervention using only quality adapters as input representations.
fusion-all: we train a fusion on the task of moderation intervention using all quality adapters and the adapter for moderation.
roberta-full: we report the result of Falk et al. (2021) who predict interventions on the same data split with full fine-tuning RoBERTa.

We use the same hyperparameters for the fusion and the single-task adapter as in experiment 2.

Can AQ adapters be applied to predicting moderator interventions? Table 4 shows F1 for interventions and F1 macro as average over the 5 splits. We consider F1 for interventions to be more important because it represents the minority class and only the positive instances are suggested to a human moderator for further evaluation. Figures 5 and 7 (Appendix) show the model-to-model calculated significance values for the almost stochastic order test. All models are outperforming the baseline. The single-task adapter yields similar results to full fine-tuning, the two feature-based models with the scores for the quality dimensions yield better results for interventions, indicating that the information on the quality dimensions is useful for this task. The best results are obtained with an adapter-fusion, provided that it also includes the adapter for moderation. This indicates that the information about the quality dimensions is complementary with a data and task specific representation (moderation-ST).

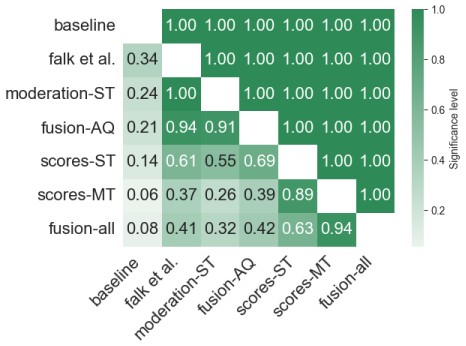


Figure 5: Almost Stochastic Order Scores (ϵ) for moderation test data for the F1 positive class, adjusted by using the Bonferroni correction. $\epsilon = 0.0$ means model in row is stochastically dominant over model in column, $\epsilon < 0.5$ denotes *almost stochastic dominance*.

Which aspects of AQ are important for predicting moderation? As discussed for the analysis of the relationship between quality dimensions in Experiment 1, an additional advantage of fusion-based models is the additional level of interpretability they provide. We investigate the relevance of each quality dimension for predicting moderation interventions using activation patterns of quality adapters. We compute the activation of each adapter of our best model (fusion-all) and visualize this as a heat-map (Fig. 4). The adapter for *impact* is the most activated. This is probably because this adapter is a good representation for distinguishing high vs. low quality comments, since the underlying dataset provides a high number of different topics (and thus can provide a good domain-independent representation). This is followed by dimensions that are important for a civic and appreciative interaction (*empathy* and *respect*) or for a solution-oriented discourse that considers the common good (*proposal*, *cgood*). The adapters for *argumentative* and *quality* add the more rational dimensions of two very different data sources, followed by the more affective and rhetorical dimensions (*story*, *narration*, *emotion*).

8 Conclusion

This work targeted the relationships between different aspects of argument and deliberative quality. We experimented with 6 datasets and 20 quality dimensions, employing adapters we learn modular representations of the targeted dimensions. We

[3]For dimensions based on binary classification we use the probability of the positive class, for the multi-class dimensions, we use the probability for each class (e.g. *common good* will be converted into three features: probability for class 1 ('no reference'), class 2 ('reference to own country') and class 3 ('reference to common good')

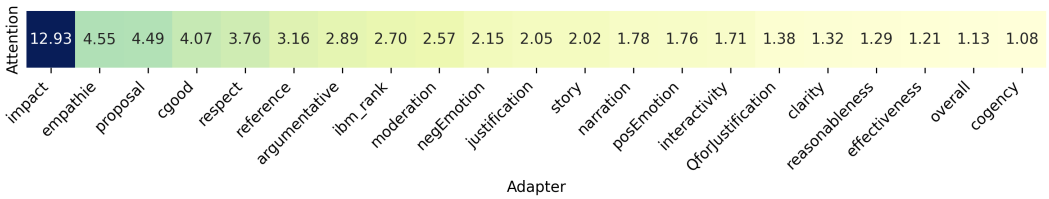


Figure 4: Predicting moderator intervention: activation for each dimension (avg. over all test instances, sum over all layers)

show that adapter fusion improves predictions in 8 dimensions out of 20. We then use the learnt adapters in the task of predicting moderator interventions - we show that information about different argument quality dimensions helps to improve the performance. Having more insights about which aspects of argument quality were activated more or less when the model that a user contribution should be moderated could help human moderators decide which aspects to focus on and information about why that model fired can increase human moderators awareness of and (ideally) confidence in automated support methods. We make models for single-task adapters and fusions and the code to train and test them available: https://github.com/Blubberli/ArgQualityAdapters.git.

9 Limitations

The datasets were created with very different motivations, the annotations were partly created by experts, partly via crowd-sourcing. The definitions of the different aspects of argument quality are also based on different theories or merely on human intuition. This work is only a first step to collect the existing data, to use it and to gain first insights about overlaps between relations based on empirical experiments. A deeper analysis of the underlying annotations and definitions is an urgent next step. Another limitation is that we compare the benefits of adapter-fusion to single-task adapters in a low-resource scenario. Because we are dealing with a large amount of different dimensions (20) additional experiments that compare this approach to full fine-tuning or traditional MLT-learning were not feasible in this work but can be conducted in the future, potentially on a smaller set of selected dimensions. On top of that we do not try to improve the state-of-the-art results for each quality dimension for each dataset. This is for the following reasons: the main focus of this work is to investigate whether adapter-fusion improves the results compared to single-task adapters, not which model works best for which data set. The SOTA results for individual dimensions in our case are either not available (Social Science datasets) or based on data-specific optimizations of the hyperparameters / architectures. We focus on a variety of dimensions and datasets, especially those coming from the social sciences. In addition to the potential improvements in results through MLT with adapter-fusion, we see the advantage above all in the modular design (depending on the annotation from future datasets, dimensions can simply be added or omitted) and the insights we can gain about the contribution of individual dimensions through attention patterns. The models in this work were partially trained on small datasets. It is necessary to investigate to what extent the models are applicable to other domains. Also the influence of the topics in the discussions (topic bias) should be investigated.

Potential Negative Societal Impacts The automatic modeling of Argument Quality bears the danger that what is considered as "high quality arguments" will be closely related to what is represented as high quality in the existing datasets. This might disadvantage certain styles of argumentation but also certain opinions that are so far underrepresented in the data. It is therefor necessary to investigate how these models behave with data with such underrepresented styles and opinions and to create new datasets with AQ with greater diversity.

Acknowledgments

We would like to thank Anne Lauscher and Agnieszka Faleńska who provided valuable feedback at various points of this work. This research has been funded by Bundesministerium für Bildung und Forschung (BMBF) through the project E-DELIB (Powering up e-deliberation: towards AI-supported moderation).

References

Milad Alshomary, Timon Gurcke, Shahbaz Syed, Philipp Heinisch, Maximilian Spliethöver, Philipp Cimiano, Martin Potthast, and Henning Wachsmuth. 2021. Key point analysis via contrastive learning and extractive argument summarization. In *Proceedings of the 8th Workshop on Argument Mining*, pages 184–189, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Eustasio Del Barrio, Juan A Cuesta-Albertos, and Carlos Matrán. 2018. An optimal transportation approach for assessing almost stochastic order. In *The Mathematics of the Uncertain*, pages 33–44. Springer.

Rotem Dror, Segev Shlomov, and Roi Reichart. 2019. Deep dominance - how to properly compare deep neural models. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2773–2785. 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.

Katharina Esau. 2022. *Kommunikationsformen und Deliberationsdynamik*. Nomos Verlagsgesellschaft mbH & Co. KG.

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.

Michael Fromm, Max Berrendorf, Johanna Reiml, Isabelle Mayerhofer, Siddharth Bhargava, Evgeniy Faerman, and Thomas Seidl. 2022. Towards a holistic view on argument quality prediction. *CoRR*, abs/2205.09803.

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

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

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

Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jiawei Low, Lidong Bing, and Luo Si. 2021. On the effectiveness of adapter-based tuning for pretrained language model adaptation. 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 2208–2222, Online. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *ICML*.

Anne Lauscher, Olga Majewska, Leonardo F. R. Ribeiro, Iryna Gurevych, Nikolai Rozanov, and Goran Glavaš. 2020a. Common sense or world knowledge? investigating adapter-based knowledge injection into pretrained transformers. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 43–49, Online. Association for Computational Linguistics.

Anne Lauscher, Lily Ng, Courtney Napoles, and Joel Tetreault. 2020b. 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.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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.

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*, dg.o '12, page 173–182, New York, NY, USA. Association for Computing Machinery.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

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

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226, Prague, Czech Republic. Association for Computational Linguistics.

Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. 2022. deep-significance-easy and meaningful statistical significance testing in the age of neural networks. *arXiv preprint arXiv:2204.06815*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

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

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

Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020. A corpus for argumentative writing support in German. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 856–869, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Appendix

A Implementation details

For training the adapters and the adapter-fusion models we use the `adapter-transformers` library (Pfeiffer et al., 2020) with `roberta-base` as a backbone. We use the default hyperparameters (learning rate of 0.0001 which was found to work empirically best in most setups (Pfeiffer et al., 2020, 2021) and a reduction factor of 16). As a maximum sequence length we use 256 which is higher than all means of the 6 datasets. We train the adapters for a maximum of 40 epochs for classification and 25 epochs for regression and use the model with the best performance (lowest MSE or highest F1 macro) on the validation set. For the adapter-fusion we also rely on the best learning rate according to Pfeiffer et al. (2021), which is $5e-5$. We lower the maximum number of epochs (25 for classification and 15 for regression). We train all models on 3 GPUs (*NVIDIA RTX A6000, each GPU has 49GB, CUDA Version 11.7*) with a batch size of 16, each model is trained with 3 seeds (5, 42, 108) for the experiments reported in section 4. We use the the adapters of one seed (42) to generate the predictions and the single-task adapters trained with that seed for the AdapterFusion in the experiment in section 7. The largest model is the AdapterFusion with 21 adapters (all quality dimensions and moderation). The training run time for this is 15.349 samples per second and 44.282 samples per second during inference. In experiment 3, for the logistic regression classifiers, we find the best hyperparameters using grid search and 3-fold cross-validation on a separate data split (L2 penalty, class weights and C=0.1).

B Datasets

The tables in the end of this Appendix (Table 13 for THF/BK, Table 14 for Europolis, Table 15 for GAQ and Table 16 for SwanRank, IBM-Rank-30k and Kialo) illustrate examples of each dataset, each example exhibits are high score (or label) of a different dimension of AQ.

Parts of the transcriptions of the Europolis dataset were not in English and automatically translated using DeepL (`https://www.deepl.com/translator`).Similarly, the online-comments from THF/BK are originally German and have been automatically translated

using DeepL. Samples of the automatic translations were verified by native speakers.

Data splits for Experiment 1: Table 5 shows the amount of training / development and test data for each corpus.

	train	dev	test
THF/BK	788	198	247
Europolis	546	140	175
Kialo	650	150	200
GAQ	650	150	200
SwanRank	650	150	200
IBM-Rank-30k	650	150	200

Table 5: Amount of train, validation and test data for each dataset. The amount for *Kialo,GAQ,SwanRank,IBM-Rank-30k* has been down-sampled to 1000 instances.

Table 6 gives an overview the positive amount of instances for each quality dimension in the training data. Most of the dimensions (except argumentative) are the minority class.

dimension	relative amount in train
posEmotion	13 %
proposal	38 %
narration	31 %
reference	41 %
argumentative	75 %
negEmotion	21 %
empathy	11 %
Qjustification	20 %

Table 6: Relative amount of positive instances for each quality dimension in the THF/BK training set.

Table 7 and 8 show the distribution of each class label for the dimensions in Europolis and the one in Kialo.

Table 9 and 10 show the mean and standard deviation for the point-wise quality scores in the training data of GAQ, SwanRank and IBM-Rank-30k.

C Experiment 1

Heuristic The following describes more details about the heuristic used to select source tasks for the multi-task experiment in section 4. To generate predictions we first train single-task adapters on the original datasets. We use the original train/val/test split for IBM-Rank-30k (train=20974, val=3208, test=6315=, GAQ (train=2746, val=1177, test=538) and Kialo (train=5170, val=1108 test=1108) and create our own split for SwanRank (train=3440, val=860, test=1075), THF/BK and Europolis (splits in Table 5.

Dimension and labels	amount	Dimension and labels	amount
interactivity		**respect**	
negative reference	**41 %**	disrespectful	**10 %**
no reference	**4 %**	implicit respect	**75 %**
neutral reference	**35 %**	explicit respect	**15 %**
positive reference others	**20 %**		
cGood		**justification**	
no reference	9 %	no justification	16 %
own country	76 %	inferior justification	40 %
common good	15 %	qualified justification	34 %
		sophisticated	10 %
storytelling			
storytelling	33 %		
no storytelling	67 %		

Table 7: Distribution of class labels for each dimension in the Europolis training set.

impact labels	relative amount in train
not impactful	**22 %**
medium impactful	**23 %**
impactful	**55 %**

Table 8: Distribution of class labels for *impact* in the kialo training set.

Table 11 shows the results of each single-task adapter on the original-sized dataset. We report the mean and standard deviation across 3 seeds.

We then take the adapter for each dimension and generate predictions for all other datasets. For feasibility, we sample 3000 instances for Kialo, IBM-rank-30k and SwanRank to generate predictions on these subsets. Based on the predictions we compute the pair-wise Spearman correlations between the AQ dimensions for each dataset. For binary classes we use the probability of the positive class as a continuous score, for dimensions with 3 to 4 classes we convert the predicted class labels into scores on a linear scale, e.g. *impact* has 3 class

dimension	mean	std
cogency	3.29	**0.65**
effectiveness	3.13	0.76
reasonableness	3.05	0.72
overall	**3.14**	0.72

Table 9: Mean and standard deviation of point-wise quality for each dimension in the GAQ corpus.

dimension	mean	std
quality	0.79	0.20
clarity	0.53	0.24

Table 10: Mean and standard deviation of point-wise quality for *clarity* and *quality* in the corresponding training sets.

dimension	performance
Pearson correlation	
overall	0.56±.00
cogency	0.54±.01
effectiveness	0.59±.01
reasonableness	0.49±.00
quality	0.55±.00
clarity	0.73±.01
F1 macro	
justification	0.46±.04
interactivity	0.35±.05
respect	0.50±.04
cgood	0.60±.04
story	0.75±.02
Q(uestion) for justification	0.89±.01
reference	0.80±.01
argument	0.77±.01
narration	0.76±.02
proposal	0.79±.01
negEmotion	0.70±.01
posEmotion	0.64±.03
empathy	0.69±.04
impact	0.52±.01

Table 11: Results on the test set for each quality dimension. Performance with standard deviation, averaged over 3 seeds.

labels: low impact, medium impact, high impact which we convert into 1, 2 and 3 respectively. We compute the pair-wise correlations for each dataset (we take the gold annotations when available) and average them them across datasets. Figure 6 shows the pair-wise correlations as a correlation matrix.

Next we sample source tasks for each target task based on the correlations. Taking different samples of dimensions we compute a threshold based on absolute correlation values and add a dimension as source task if the correlation to the target dimension exceeds the computed threshold. We consider the following setups:

fusion corr ALL We select the source tasks from all dimensions if the absolute correlation value is higher than 0.24 (corresponds to the third quartile of all correlations).

fusion corr DQ We consider only source tasks from datasets with a deliberative focus (THF/BK, Europolis, Kialo). The tasks are sampled from 14 dimensions and the threshold is 0.15 (third quartile of all correlations between the 14 dimensions).

fusion corr AQ We extract the source tasks from all dimensions that stem from more general argumentative contexts (IBM-rank-30-k, GAQ, swanson). The threshold is based on the correlations between the 6 dimensions (0.54, the second quartile due to the high correlations between the GAQ dimensions).

Table 12 shows the source task dimensions for each target dimension, depending on the setup (fusion corr ALL, fusion corr DQ, fusion corr AQ). Each dimension is learned using between 1 and 9 other dimensions as source tasks. For *respect* there were no dimensions with a high enough correlation in any of the setups.

D Experiment 2: predicting moderator interventions.

Figure 7 shows the significance matrix between all models for the task of predicting moderator interventions for the F1 macro score.

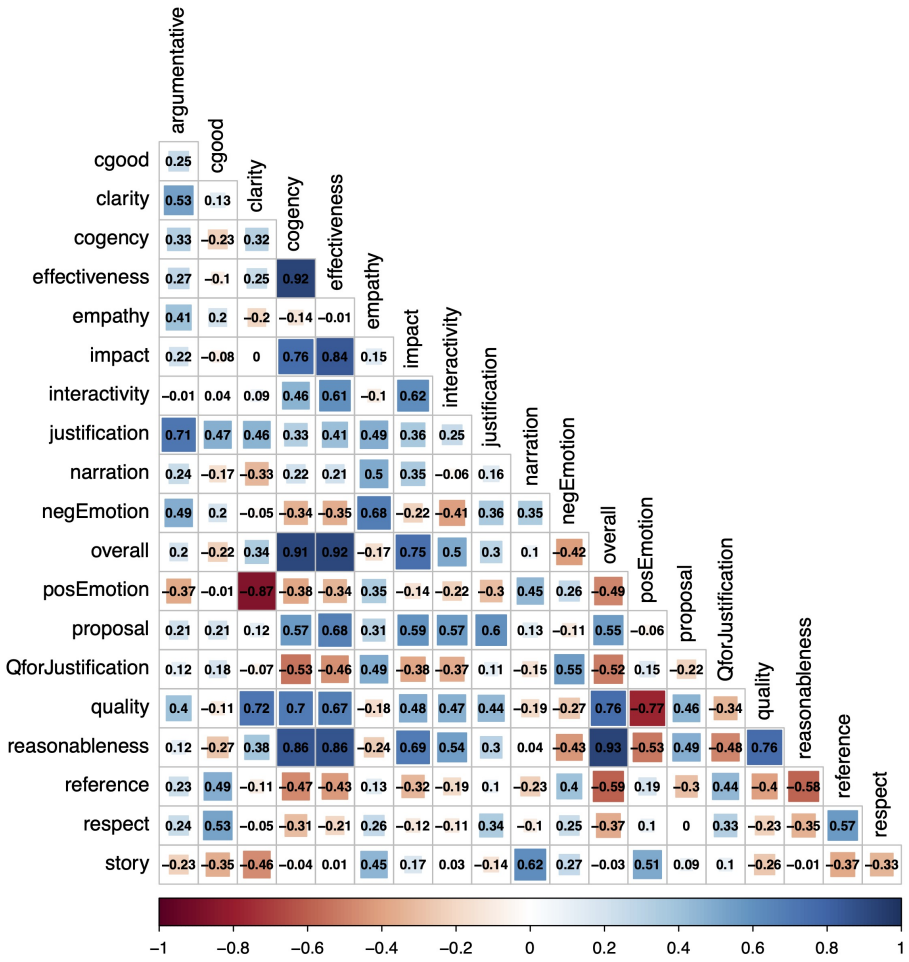
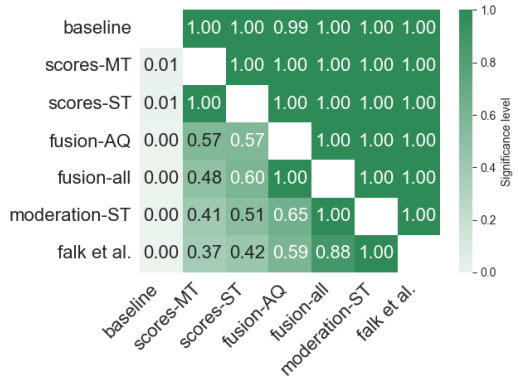


Figure 6: Pairwise Spearman correlations between all quality dimensions based on single-task adapter predictions. Average across all 6 datasets.

target task	additional source tasks using all dimensions	additional source tasks using only dimensions from deliberative context	additional source tasks using only dimensions from general argumentation
overall	impact, effectiveness, proposal, quality, reference, reasonableness, clarity, cogency, justification	-	effectiveness, reasonableness, cogency
cogency	impact, effectiveness, overall, quality, reference, reasonableness, clarity, justification	-	effectiveness, overall, quality, reasonableness
reasonableness	impact, effectiveness, overall, quality, reference, clarity, cogency, justification	-	effectiveness, overall, cogency
effectiveness	impact, proposal, overall, quality, reference, reasonableness, clarity, cogency, justification	-	overall, quality, reasonableness, cogency
quality	effectiveness, overall, posEmotion, reasonableness, clarity, argumentative, cogency, justification	-	effectiveness, cogency
clarity	effectiveness, overall, posEmotion, quality, reasonableness, story, argumentative, cogency, justification		-
justification	effectiveness, proposal, overall, quality, negEmotion, reasonableness, clarity, argumentative, cogency	proposal, empathy, negEmotion, cgood, argumentative	-
story	empathy, reference, clarity, narration	empathy, posEmotion, cgood, reference, narration	-
interactivity	-	negEmotion	-
cgood	-	story, justification	-
posEmotion	quality, clarity, narration	story, argumentative, narration	-
proposal	effectiveness, overall, reference, justification	reference, justification	-
narration	empathy, posEmotion, negEmotion, story	empathy, posEmotion, negEmotion, reference, story, argumentative	-
reference	effectiveness, proposal, overall, reasonableness, story, cogency	proposal, story, narration	-
argumentative	quality, negEmotion, clarity, justification	empathy, posEmotion, negEmotion, narration, justification	-
negEmotion	empathy, argumentative, narration, justification	empathy, interactivity, argumentative, narration, justification, QforJustification	-
empathy	negEmotion, story, narration	negEmotion, story, argumentative, narration, justification, QforJustification	-
Qjustification	-	empathy, negEmotion	-
impact	overall, reasonableness, cogency, effectiveness	-	-

Table 12: Multi-task experiments: target dimension with source dimensions used as input adapters for adapter-fusion.



(a) F1 macro

Figure 7: Almost Stochastic Order Scores (ϵ) for moderation test data for the F1 macro score, adjusted by using the Bonferroni correction. $\epsilon = 0.0$ is means model in row is stochastically dominant over model in column, $\epsilon < 0.5$ denotes *almost stochastic dominance*.

example	dimension
On the one hand, our lignite is needed to maintain an affordable and reliable energy supply (and based on physical and economic laws, will still be needed in 50 years) and on the other hand, our lignite can do more than just be burned to generate electricity.	argumentative
In New Zealand, residents of the Pacific island of Tuvalu have already been granted the right to asylum - on the grounds of climate change. Who is asking for their recovery? How do people who have been forced to flee their homes due to the global burning of fossil fuels and the resulting DECREASED global warming read our news and debates? "Act only according to that maxim by which you can at the same time will that it become a general law." If we include Immanuel Kant's thoughts in the guiding decision, shouldn't lignite mining really end at the A61 and RWE workers be supported in the corporation's structural transformation in a way that provides well for them and their families?	empathy
Classical music for all Once a week the Berlin Philharmonic Orchestra should play at THF for ALL Berliners. This way even families with little money can enjoy classical music. The prices at the Philharmonie or concerts by other great musicians are so immensely high that only higher earners can afford it. This is an outrage because they are subsidized by us and we can't even afford to go.	proposal
I think #person is more than right and I share his opinion... Lignite has and should continue to have a place here in the region. Good luck	reference (to other discourse participants)
I have been to Holzweiler many times. The experiences from Immerath and Borschemich show that the club life in an intact village does not suffer due to the resettlement. On the contrary, it strengthens the feeling of togetherness and allows the clubs to flourish.	narration
But I also think what #person wrote is great. One notices from it that not immediately a rejection against it prevails but rather a certain concern. In particular here around animals. You also notice that there is still a great ignorance. I find great that you have expressed yourself. I think the discussions here should be there to reduce possible worries and prejudices. Thank you	positive emotion
I have been following what has happened to lignite for many years and I think it is terrible. I've lived in the Rhineland for years and it's easy to live with the changes caused by lignite. More and more good jobs are disappearing in Germany. My last employer is already cutting well-paid jobs due to the low oil price.	negative emotion

Table 13: Examples for each Quality Dimension in THF/BK

example	dimension
I have friends from Latin America and many other places and they work and they pay in a pot. So, I'm a mother and well, unfortunately, if I have to go to another country, well, I try to integrate in the country I'm going to. I'm not going to go there, to impose my goals [?], my way of seeing, no. I'm going there to work and not to steal. And there is something else. But again, I'm holding back.	storytelling
I don't know if you can regulate it well, how many people immigrate or emigrate or whatever. I think it's important to create a basis for all people to be able to live in their country. Because I think that is actually the main cause. That many industrialized countries are bleeding small countries or poor countries dry and taking away their livelihood. And that's why people emigrate, because they no longer have anything to eat, because they can no longer find work in their country, and because life in the industrialized countries is simply made out to be nicer or better. In order to be able to ultimately prevent an immigration policy, illegals, I believe that you first have to change the basis in the other countries, that is, the countries of origin. Create a basis. Life base.	sophisticated justification
Well, I am of the opinion that simply in the population the term EU is seen completely wrong; one always wants only something and one wants to give nothing. I am simply of the opinion that it should be a community and a community simply has to support the weaker ones and the stronger ones simply have to give. I think this is the basic problem of the EU and I think it's very nice that today and in the next few days this could contribute to the fact that this spirit, which was really brought into being by Robert Schuman and by all those who have worked so hard for the EU, could be recognized and a community could really take place; at least in the microcosm now.	reference to common good
Here, when we are talking about immigration, it should be first identified why a certain person left his country. Just like my friend before said to be a refugee is also a man, which probably feels bad in his own country. For example, when his country's situation does not give him a life in dignity. So I think that every country should identify immigrants and help them in certain ways, for example with social benefits. I know that some countries for example Poland are not rich countries, so they need EU help in such a matter. Especially the countries where immigration is quite high.	explicit respect
Yes I completely agree with what this gentleman just said because I think we have created ghettos, we have - at the moment - people who live very very badly, immigrants who live very very badly, who are already unemployed, who have enormous problems of integration and I think we should already make an effort to integrate these people who are well in our countries and then we see what we can do to bring in others, we must already take care of the people who are on our territory and who are living very badly and who are unemployed, who are poorly cared for, who have problems with their children, school problems, problems with papers and I think that we must already arrive once we have properly resolved these problems and that we will have sooner than bring people in and make them unhappy - I think that it is perhaps worse than doing something more moderate.	positive reference to other discourse participant
positive reference	

Table 14: Examples for each Quality Dimension in Europolis

example	dimension
I'm a fairly tolerant human being and am in no way an advocate of the death penalty. I also understand that a short sentence and rehabilitation is also an effective form of justice in terms of re-offending in many countries. However I think there's gotta be a line drawn somewhere in which a person entirely loses their liberty and autonomy if the crime they committed was as heinous as the one committed by Mr Breivik. Also, even though I am aware that there is a good chance he will still remain behind bars for the rest of his life, the possibility that he won't baffles and worries me. Please CMV.	high cogency (5)
Well, this topic has raised lots of questions lately particularly in France. This is where I stand: - Wearing a burqa should be a matter of choice, just as women choose to wear anything else, regardless of any religious manifestations. - Wearing the burqa shouldn't be banned, and shouldn't be forced to women either; it should be a personal choice. - When talking about choices, it's the society that gives these choices according to what the majority thinks, hence the more civilized and democratic a society is the more choices people have. - It's basically a matter of respect, if a woman chooses to wear it then we should respect that, we can't force her not to wear it, as we can't force her to wear it: free will :) , on the other hand that woman should respect and obey all the security issues that comes along with wearing it.	high effectiveness (5)
The point of daylight savings is to make our numeric time cycle fit with the Sun's time cycle. In other words, standardize the time of day in which the sun is shining. This way, people and businesses can keep their operating hours steady without working in the dark, and less electricity is used. Most arguments I've heard against it pertain to the inconvenience of changing clocks and accounting for gained/lost hour, but with most clocks being digital and synced up to DST nowadays, that's becoming less and less of a problem. And besides, one day of inconvenience in exchange for a whole season of "correct" daylight seems like a pretty good deal to me.	high reasonableness (5)
I believe property is a social construct that is only justified through appeals to utility. In other words, any particular set of property laws are only justified insofar as they make people better off, in terms of their capabilities. Most Libertarians I've debated with either believe property rights are somehow fundamental(natural or God-given) or develop out of other moral principles, like the NAP. The first option appeals to non-existent entities. The second is circular, as what NAPer's define as aggression is violation of property rights, and violations of property rights is defined in terms of the NAP.	high overall (5)

Table 15: Examples for each Quality Dimension in GAQ

example	dimension	Dataset
A basic principle of punishment is that it should be proportional to the crime, and therefore capital punishment is the only legitimate response to a crime such as first degree murder.	high quality (1.0)	IBM-Rank-30-k
When voters are able to make an impact and change their votes more often they will feel more engaged with the political process, and get more involved in politics.	high impact	Kialo
First a prediction is made from an hypothesis of some observation that must be true if the hypothesis is correct.	high clarity (1.0)	SwanRank

Table 16: Examples for each Quality Dimension in SwanRank, Kialo and IBM-rank-30-k