An empirical analysis of diversity in argument summarization

Meer, M.T. van der; Vossen, P.T.J.M.; Jonker, C.M.; Murukannaiah, P.K.; Graham, Y.; Purver, M.

An Empirical Analysis of Diversity in Argument Summarization

Michiel van der Meer
LIACS
Leiden University
m.t.van.der.meer@liacs.leidenuniv.nl

Piek Vossen
CLTL
Vrije Universiteit Amsterdam
p.t.j.m.vossen@vu.nl

Catholijn M. Jonker
Interactive Intelligence
TU Delft
c.m.jonker@tudelft.nl

Pradeep K. Murukannaiah
Interactive Intelligence
TU Delft
p.k.murukannaiah@tudelft.nl

Abstract

Presenting high-level arguments is a crucial task for fostering participation in online societal discussions. Current argument summarization approaches miss an important facet of this task—capturing *diversity*—which is important for accommodating multiple perspectives. We introduce three aspects of diversity: those of opinions, annotators, and sources. We evaluate approaches to a popular argument summarization task called Key Point Analysis, which shows how these approaches struggle to (1) represent arguments shared by few people, (2) deal with data from various sources, and (3) align with subjectivity in human-provided annotations. We find that both general-purpose LLMs and dedicated KPA models exhibit this behavior, but have complementary strengths. Further, we observe that diversification of training data may ameliorate generalization. Addressing diversity in argument summarization requires a mix of strategies to deal with subjectivity.

1 Introduction

Getting an overview of the arguments concerning controversial issues is often difficult for those participating in ongoing discussions. This is because there are many points being communicated, no way to track which arguments were already encountered, and haphazard miscommunication or conflict. Automatic summarization is a way to provide a comprehensible overview of the opinions (Nenkova et al., 2011; Angelidis and Lapata, 2018). However, generating summaries representative of the arguments involved in a discussion is difficult (Bar-Haim et al., 2020a). Argument summarization extends beyond text summarization because it separates argumentative and non-argumentative content, preserves the argumentative structure, and provides explicit stances on a central claim or hypothesis.

Summarizing arguments is challenging in many contexts, but the potential impact is high. For instance, after summarizing the arguments from societal discussions, the extracted arguments may shape new policies and be used to justify decision-making (Arana-Catania et al., 2021; Gürkan et al., 2010). Similarly, businesses depend on review data to find customer feedback, which can be used to steer product design (Archak et al., 2007).

Although arguments are often summarized by hand in practice (e.g., Mouter et al., 2021; McLaren et al., 2016; Nahm, 2013), recent developments in Argument Mining (AM) allow automatic analysis of argumentative text (Lawrence and Reed, 2020). Obtaining summaries that faithfully represent open-ended opinions requires careful evaluation, especially in sensitive contexts, e.g., summarizing citizen feedback (Egan et al., 2016; Misra et al., 2015).

One approach for generating comprehensive summaries of arguments is Key Point Analysis (KPA, Bar-Haim et al., 2020a). In KPA, a corpus of opinions is analyzed for the *key points*, those arguments that are salient and repeated multiple times. However, some aspects of the KPA experimental design misalign with respect to real-world applications. We illustrate these blind spots, in particular, when applied to summarizing online societal discussions. We highlight three dimensions of **diversity** that are central to empowering citizens' opinions at scale (Shortall et al., 2022): (1) incorporating the long tail of opinions, (2) being robust in handling data from multiple sources, and (3) including diverse perspectives from annotators.

How current KPA approaches deal with the above dimensions of diversity is unexplored. We conduct an empirical study of different argument summarization approaches by incorporating the standardized benchmark and two other datasets to experiment. We develop specific analyses to un-

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2028–2045

cover how KPA approaches fare on each dimension of diversity. In addition to the existing approaches, we use LLMs by prompting them to perform KPA, as they may be an attractive alternative to current models.

Applying KPA approaches across several datasets that vary in how they address diversity leads to mixed results. KPA performance degrades when dealing with low-frequency opinions, i.e., opinions repeated by relatively few individuals. Further, we observe that KPA approaches disregard subjective interpretations among individual annotators. Finally, they generalize poorly across data sources when used in transfer learning settings, though approaches reveal complementary merits across tasks.

Contributions (1) We critically examine three dimensions of diversity—of opinions, sources, and annotators—in the KPA setup. (2) We analyze the behavior of existing metrics on one existing and two newly adapted datasets. (3) We analyze multiple methods, including prompt-based LLMs, broadening the scope of methods that can perform KPA.

2 Related Work

We outline three lines of related work: key point analysis, opinion summarization, and diversity.

2.1 Key Point Analysis

KPA serves to separate argumentative from non-argumentative content, and condense argumentative content by matching arguments to key points (Bar-Haim et al., 2020a). Key points can be seen as high-level arguments that capture the gist of a set of arguments. While most work on KPA selects high-quality arguments as representatives, generating novel key points has been proposed as an alternative (Syed et al., 2021). KPA has been applied across topics using data from discussion portals or online reviews (Bar-Haim et al., 2020b, 2021a). KPA is usually divided into Key Point Generation and Key Point Matching steps (see Section 3.1).

Multiple approaches exist for KPA (Friedman-Melamed et al., 2021). Modeling choices consist of popular Transformer models such as BERT (Phan et al., 2021), enhanced representational quality using contrastive learning (Alshomary et al., 2021), and the incorporation of clustering techniques (Li et al., 2023). Our work aims to investigate some of the modeling choices employed in these works. For instance, in Li et al. (2023), the authors discarded unmapped arguments, which may hurt the ability the represent minority opinions.

2.2 Opinion Summarization

Opinion summarization aims to generate summaries of an individual's subjective opinions (Inouye and Kalita, 2011; Bhatia, 2020), often applied to product reviews (Chu and Liu, 2019). Leveraging Transformer models is popular for opinion summarization (Angelidis et al., 2021; Amplayo et al., 2021), though generic extractive summarization techniques are strong baselines (Suhara et al., 2020). Measuring bias in generated summaries has seen recent interest, specifically acknowledging that diverse opinions should be taken into account (Huang et al., 2023; Siledar et al., 2023) or postulating that diversity is a desirable trait when generating opinions (Alshomary et al., 2022; Wang and Ling, 2016) . Our work applies these techniques to argumentation to obtain a high-level summary of opinions, and analyses differences in behavior for (in-)frequent viewpoints.

2.3 Diversity in Societal Decision Making

Sensitive decision-making contexts call for responses rooted in reason that serve social good rather than specific interests. One way of obtaining such responses is through evidence-based policymaking, which involves stakeholders and the broader public to strike decisions (Cairney, 2016). Citizen participation improves the support of the decisions when some requirements are met (Mansbridge et al., 2012). A key factor among those requirements is the involvement of a diverse group of citizens, independently voicing opinions (Surowiecki, 2005). Approaches to summarizing arguments in such citizen feedback face similar requirements.

In Argument Mining, we find recent work that aligns with these views, e.g., by a strong focus on the diverging perspectives among annotators in AM tasks (Romberg, 2022). Further, some preliminary work adjusts visualization for minority opinions (Baumer et al., 2022). However, in terms of data sources, most work is still centered on English-speaking content, with few multi-lingual or multi-cultural resources available (Vecchi et al., 2021).

3 Method

We formulate the KPA subtasks—*Key Point Generation* (KPG) and *Key Point Matching* (KPM). We then introduce the three dimensions of diversity and consider them when applying KPA.

3.1 Task setup

We outline the two subtasks that constitute KPA, as originally introduced by (Friedman-Melamed et al., 2021).

Key Point Generation (KPG) focuses on generating *key points* $\mathcal{K}$ given a corpus of arguments $\mathcal{D}$ on a particular claim. Key points are high-level arguments that capture the gist of a collection of arguments. Key points oppose or support the claim.

Key Point Matching (KPM) *matches* arguments to key points. An argument matches a key point if the key point directly summarizes the argument, or if the key point represents the essence of the argument. We ensure that the stance of the key point (pro or con) matches the stance of the argument. Formally, given a set of key points $\mathcal{K}$ and a corpus $\mathcal{D}$, we score the match between an argument $d \in \mathcal{D}$ and a key point $k \in \mathcal{K}$ with a matching model $M(d, k)$. Assigning arguments to key points using match scores is flexible, and multiple strategies can be taken to reach a final decision (e.g. imposing a match score threshold) (Bar-Haim et al., 2020b). Since the assignment strategy is largely context-dependent, we evaluate the scoring mechanism itself, instead.

3.2 Modeling Diversity in Key Point Analysis

We focus on three main aspects of diversity.

(1) Long tail opinions Several NLP models imitate biases that exist in datasets (Blodgett et al., 2020). For argument summarization, a form of bias is focusing on majority arguments, leading to possible misrepresentations. Failing to capture low-frequency arguments runs the danger of further estranging underrepresented viewpoints (Klein, 2012). These methods need active correction from humans to account for this "long tail of opinions" (van der Meer et al., 2022). For the KPA task, approaches have largely unknown behavior on capturing the long tail of opinions (Mustafaraj et al., 2011). Additionally, LLMs struggle with learning long-tail knowledge (Kandpal et al., 2023), aggravating this issue. We experiment with subsampling the datasets to investigate the imbalanced data settings, which are representative of real-world use cases.

(2) Annotators Datasets are labeled using a mix of crowd and expert annotators. Querying experts for key points may leave the impacted users (e.g., lay citizens) out of consideration (Cabitza et al., 2023). Similarly, labels stemming from crowd annotation that are filtered for high agreement may disregard controversial or diverse opinions. Disagreement is a complex signal that includes subjective views, task understanding, and annotator behavior (Aroyo and Welty, 2015). Having access to non-aggregated annotations would, e.g., allow for further modeling of patterns (Davani et al., 2022) or the reasons (Liscio et al., 2023) underlying opinions. We investigate whether models trained on such annotations can identify disagreement.

(3) Data sources Existing works investigate cross-domain generalization of KPA methods using data stemming from 1a single dataset, focusing on a cross-topic setting (Bar-Haim et al., 2020b; Samin et al., 2022; Li et al., 2023). This dataset is gathered at a specific time. As discussions evolve, more nuanced positions may become relevant, and new real-world events impact the opinions. Further, these discussions usually take place on a single platform (e.g., Reddit threads, Twitter discussions), inheriting biases from the source (Hovy and Prabhumoye, 2021). Measuring the performance of KPA approaches should rely on diverse datasets, based on data gathered from different sources at different points in time. There have been some efforts in applying KPA across different contexts (Gretz et al., 2023; Bar-Haim et al., 2021a; Cattan et al., 2023), but they apply approaches to a single dataset at a time, making direct comparison difficult. Our work examines the cross-dataset performance of these approaches to assess their relative strengths and weaknesses.

Table 1 shows the current datasets, and how they relate to the dimensions discussed above. In all three datasets, the arguments stem from user-submitted content. In one dataset, low-frequency arguments (i.e., opinions repeated by few individuals) are disregarded. Further, the ArgKP benchmark relies on expert-generated key points and does not include annotator-specific match labels. Perspectrum contains aggregated counts of match labels, but due to aggregation, we can-

Dataset	Data Source	Filter low freq.	Key Points source	Non-aggregated annotation	IRR
ArgKP	Human annotation	✓	Expert	✗	0.50-0.82 (κ)
PVE	Citizen consultation	✗	Crowd	✓	0.35 ($\kappa^{\dagger}$)
Perspectrum	Debate platforms	✗	Crowd	✗	0.61 (κ)

Table 1: Datasets and their diversity characteristics when considering the KPA task. The inter-rater reliability (IRR) is measured via Cohen's κ scores or prevalence and bias-adjusted Cohen's $\kappa^{\dagger}$ (PABAK, Sim and Wright, 2005).

not identify annotator-specific patterns. Lastly, the inter-rater reliability differs for each dataset, with wide ranges, showing that the tasks are fundamentally subjective. We employ these three datasets for evaluating various KPA approaches and dive deeper into the three aspects of diversity.

4 Experimental Setup

We describe the data, KPA methods, and metrics involved in our experiments. We make our source code[1] and data (van der Meer et al., 2024) publicly available.

4.1 Data

Most work on KPA has used the dataset introduced by Friedman-Melamed et al. (2021) in a shared task. We add two new datasets that match the KPA subtasks but have different characteristics.

ArgKP We adopt the shared task dataset, keeping the same split across claims as the original data. The ArgKP dataset contains claims taken from an online debate platform, together with crowd-generated arguments and expert-generated key points (Bar-Haim et al., 2020a). The arguments were produced by asking humans to argue for and against a claim, followed by filtering on high-quality and clear-polarity arguments. Key points were generated by an expert debater, who generated the key points without having access to the arguments. The final test set was collected after the initial dataset and has been curated to match some of the distributional properties of the training and validation sets.

PVE We use the crowd-annotated data stemming from a human-AI hybrid key argument analysis (van der Meer et al., 2022) based on a Participatory Value Evaluation (PVE), a type of citizen consultation. In this consultation process, citizens were asked to motivate their choices for new COVID-19 policy through text, which formed a set of comments for each proposed policy option. The performed key argument analysis resulted in crowd-generated key points, matching individual comments to key points per option. Since this is a small dataset, we only use it for evaluation.

Dataset	Train	Val	Test
ArgKP	24 (21K)	4 (3K)	3 (3K)
PVE	–	–	3 (200)
Perspec.	525 (6K)	136 (2K)	218 (2K)

Table 2: Number of claims (and arguments) when splitting the dataset into training, validation, and test sets.

Perspectrum Similar to ArgKP, Perspectrum contains content from online debate platforms. It extracts claims, key points, and arguments from the platform directly (Chen et al., 2019). Part of the dataset is further enhanced by crowdsourcing paraphrased arguments and key points. The Perspectrum dataset is ordered into claims, which are argued for or against by perspectives, with evidence statements backing up each perspective. We use perspectives as key points, and evidence as arguments. We retain the same split over claims as the original data. The authors provide aggregated annotations on the match between arguments and key points. While this allows us to compute the agreement scores per sample, we cannot distil individual annotator patterns.

4.2 Approaches

We investigate different approaches with respect to their performance on the aspects of diversity. Appendix A includes a detailed overview of the setup, parameters, and prompts. Similar to summarization techniques, most KPG methods are either *extractive*, taking samples as representative key points, or *abstractive*, formulating new key points as free-

[1] https://github.com/m0re4u/argument-summarization

form text generation (El-Kassas et al., 2021).

ChatGPT We use the OpenAI Python API (OpenAI, 2023) to run the KPA task by prompting ChatGPT. We differentiate between open-book and closed-book prompts. For the open-book prompts, we input the claim and a random sequence of arguments up to the maximum window (given a response size of 512 tokens) in the KPG task. For the closed-book model, we only input the claim, and the model synthesizes key points. In both approaches, KPG is abstractive. In KPM, ChatGPT predicts matches for a batch of arguments at a time, all related to the same claim.

Debater We use the Project Debater API (IBM Research, 2023), which supports multiple argument-related tasks, including KPA (Bar-Haim et al., 2021b). This approach uses a model trained on ArgKP and performs extractive KPG. We query the API for KPG and KPM separately.

SMatchToPR We adopt the approach from the winner of the shared task, which uses a state-of-the-art Transformer model and contrastive learning (Alshomary et al., 2021). During training, the model learns to embed matching arguments closer than non-matching arguments. These representations are used to construct a graph with embeddings of individual argument sentences as nodes, and the matching scores between them as edge weights. Nodes with the maximum PageRank score are selected as key points. In our experiments, the model is trained using the training set of ArgKP and PerspectrumM. This method performs extractive KPG. We experiment with RoBERTa-base and RoBERTa-large to estimate the effect of model size (Liu et al., 2019).

4.3 Evaluation Metrics

We evaluate models for KPG and KPM separately. For KPG, we adopt the set-level evaluation approach from Li et al. (2023). For KPM, we reuse the match labels provided by each dataset.

4.3.1 KPG

KPG can be considered as a language generation problem (Gatt and Krahmer, 2018) for evaluation. We rely on a mixture of reference-based and learned metrics, measuring both lexical overlap and semantic similarity. We use the following metrics:
ROUGE-(1/2/L) to measure overlap of unigrams, bigrams, and longest common subsequence, respectively. We average scores for all stance and claim combinations. Additional details on the ROUGE configuration are in Appendix A.3.
BLEURT (Sellam et al., 2020) to measure the semantic similarity between a candidate and reference key point, which correlates with human preference scores. BLEURT introduces a regression layer over contextualized representations, trained on a set of human-generated labels.
BARTScore (Yuan et al., 2021) to evaluate the summarization capabilities directly by examining key point generation. In contrast to BLEURT, BARTScore evaluates the likelihood of the generated sequence when conditioning on a source.

For each metric $\mathcal{S}$ that scores the overlap between two key points, we aggregate scores into Precision P and Recall R scores using Equations 1 and 2. For P, we take the maximum score between a generated key point a and the reference key points $\mathcal{B}$, averaging over all $n = |\mathcal{A}|$ generated key points. We perform the analogous for R. We report F_1 scores to balance precision and recall.

$$P = \frac{1}{n} \sum_{a \in \mathcal{A}} \max_{b \in \mathcal{B}} \mathcal{S}(a, b) \quad (1)$$

$$R = \frac{1}{m} \sum_{b \in \mathcal{B}} \max_{a \in \mathcal{A}} \mathcal{S}(a, b) \quad (2)$$

4.3.2 KPM

We perform the KPM evaluation by obtaining match scores for key point-argument pairs. That is, for a key point k and an argument d, we check if a new model used in the KPA method would assign d to k. We reuse existing labels and do not use the results from KPG. Since we do not consider unlabeled examples between arguments and key points, we do not need to distinguish for undecided labels (as in Friedman-Melamed et al. (2021)).

We evaluate each approach using mean average precision (mAP), taking the mean over average precision scores computed for claims C. Given a claim, we compute precision P_τ and recall R_τ for all match score thresholds τ, as in Equation 3. In case an approach outputs a binary match label instead of scores, we remap the scores to 0 and 1 for non-matching and matching pairs, respectively.

$$\text{mAP} = \sum_{C} \frac{\sum_\tau (R_\tau - R_{\tau-1}) P_\tau}{|C|} \quad (3)$$

5 Results and Discussion

First, we report on the KPG and KPM evaluation. Then, we analyze how the aspects of diversity im-

pact performance beyond a cross-dataset evaluation. We show results when conditioning on the long tail of opinions, look into the connection between annotator agreement and match score, and how performance changes for diverse data sources.

5.1 KPG Performance

Table 3 shows the results of KPG evaluation. Overall, no single approach performs best across all datasets. All models perform best on ARGKP except for closed-book ChatGPT, which performs the best on the PVE dataset. Thus, by adopting diverse datasets, we demonstrate that experimenting with a single dataset may inflate KPG performance.

ChatGPT consistently scores well on ROUGE and semantic similarity. This indicates that the abstractive generation of key points is beneficial. For PVE, we observe a strong tendency for open-book ChatGPT to adjust the generated key points to the linguistic style of the arguments. This clashes with the reference key points, which are paraphrased to make sense without the context of the original arguments. Hence, the closed-book model, which does not observe the source arguments, performs better, adopting more neutral language.

SMatchToPR performs best for PERSPECTRUM. Although general-purpose LLMs are strong in zero-shot settings, a dedicated model for representing arguments achieves state-of-the-art results. The Debater approach is ranked lowest across all datasets, showing that training on a single dataset generalizes poorly to other datasets.

ROUGE and semantic similarity scores mostly agree, except for BLEURT on ARGKP. Here, we see that SMatchToPR slightly outperforms ChatGPT. We attribute this to the optimized representational qualities of SMatchToPR: it selects key points with high semantic similarity to many arguments, which is similar to how BLEURT provides scores based on contextualized representations.

Increasing model size (of SMatchToPR) improves performance for PERSPECTRUM, but not for ARGKP and PVE. Because PVE is small, the pool of sentences to pick key point candidates from is limited, and possible improvements of the model are negligible when extracting the key points. For ARGKP, the ROUGE scores deteriorate, while the semantic similarity scores improve slightly. Intuitively, this matches expectations: the model can navigate the embedding space better, selecting key points that may be phrased differently but contain semantically similar content.

5.2 KPM Performance

Table 4 shows the results of KPM evaluation. ChatGPT, despite its strong performance on KPG, does not accurately match arguments to key points. Interestingly, the Debater outperforms the SMatchToPR model on the ARGKP dataset, but SMatchToPR is stronger on the PVE and PERSPECTRUM datasets. SMatchToPR's strong performance on PERSPECTRUM and ARGKP is expected–they were included in its training. However, its good performance on PVE is interesting and it suggests that generalization is aided by more diverse data in training.

5.3 Analysis

Long tail diversity Most key points and claims are heavily skewed in the number of data points, except for PVE. Even for ARGKP, where key points with few matching arguments were removed, there is a strong imbalance across claims and key points in terms of associated arguments (App. 3).

Following this imbalance, we sort key points by the number of associated arguments such that the least frequent key points are considered first. Then, we introduce a cutoff parameter f to include arguments from a fraction of key points, starting with the least frequent. Using this parameter we perform matching only on low-frequency key point–arguments matches. This allows us to investigate the approaches' performance in the long tail.

When we limit data usage by taking long tail arguments first, the performance of the KPA approaches, mainly on ARGKP and PERSPECTRUM, decreases as shown in Figure 1. This shows that the ability to correctly match arguments is contingent on the frequency of the arguments. In some cases, the arguments associated with key points with the fewest matches can be matched, but there is a strong performance loss for low values of f. Across all datasets, ChatGPT suffers consistently in mAP when conditioning on low-frequency key points. For SMatchToPR on PERSPECTRUM, there is almost no effect, showing that representation learning may positively impact the matching of key points to arguments even with low amounts of data.

Performing the same experiment for KPG results in similar results: key points with a low number of matched arguments are harder to represent well.

Next, we investigate whether the arguments in the long tail are different from the majority. Here, the long tail consists of arguments for key points

Dataset	Approach	R-1	R-2	R-L	BLEURT	BART
ArgKP	ChatGPT	**34.3**	**12.5**	**30.3**	0.556	**0.540**
	ChatGPT (closed book)	29.5	7.1	25.6	0.314	0.256
	Debater	25.6	5.5	22.5	0.334	0.307
	SMatchToPR (base)	31.7	11.1	29.7	0.553	0.494
	SMatchToPR (large)	30.5	8.3	26.8	**0.563**	0.497
PVE	ChatGPT	18.5	3.9	15.3	0.329	0.369
	ChatGPT (closed book)	**27.1**	**8.6**	**21.4**	**0.376**	**0.378**
	Debater	13.3	0.0	13.3	0.294	0.188
	SMatchToPR (base)	21.3	3.7	16.6	0.351	0.344
	SMatchToPR (large)	21.3	3.7	16.6	0.351	0.344
Perspectrum	ChatGPT	21.3	5.7	18.2	0.355	0.322
	ChatGPT (closed book)	17.1	3.8	15.0	0.291	0.258
	Debater	9.4	0.4	8.5	0.197	0.210
	SMatchToPR (base)	22.5	6.5	19.3	0.257	0.232
	SMatchToPR (large)	**22.7**	**6.7**	**19.4**	**0.403**	**0.363**

Table 3: ROUGE scores and semantic similarity scores for the Key Point Generation task.

	mAP		
Name	ArgKP	PVE	Perspec.
ChatGPT	0.17	0.27	0.46*
Debater	**0.82**	0.51	0.51
SMatchToPR (base)	0.76	0.53	0.80
SMatchToPR (large)	0.80	**0.61**	**0.82**

Table 4: Results for the Key Point Matching task. Closed-book ChatGPT scores are not available, since its KPA is made without observing arguments. The scores for ChatGPT on Perspectrum (*) were estimated on a subset of the test set to cut down costs.

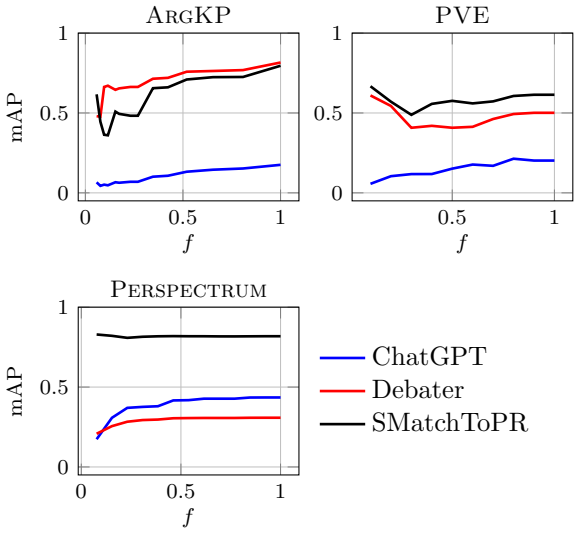

Figure 1: KPM performance when limiting data usage to a fraction f, starting with long tail first.

that see less than the median number of arguments per key point. We examine whether the sets of lexical items—noun phrase chunks (NPs) and entities—mentioned in the long tail arguments are included in the majority and vice versa. We also inspect the relative frequency of the shared lexical items via Kendall τ correlation on the NP and entity frequency rankings. Table 5 shows these results.

We see a large overlap of NPs and entities for ArgKP between the long tail and the frequent key points. We attribute this to the filtering of low-frequency data during dataset construction. For the other two datasets, we observe much less overlap—in most cases, more than half of the noun phrases and entities are unique to either part of the dataset. The only exception here is Perpectrum, where roughly 40% of the NPs and entities in the long tail are unique. When comparing the ranks of the intersecting lexical items, we observe moderate (but significant) rank correlation scores. Thus, the overlapping NPs and entities may not be in different frequencies in the two parts of the datasets. However, there is a strong indication of unique items in the long tail, in at least two of our datasets, showing that the long tail may contain novel insights.

Annotator agreement Due to subjectivity in the annotation procedures, we expect annotators to rate argument–key point matches differently. We in-

		NP		Entity			
Left	**Right**	left−right	right−left	left−right	right−left	NP-τ	Ent-τ
ArgKP-long_tail	ArgKP-majority	0.168	0.234	0.191	0.273	0.216*	0.373*
PVE-long_tail	PVE-majority	0.638	0.787	0.719	0.809	0.521*	0.389
Perspec.-long_tail	Perspec.-majority	0.397	0.807	0.401	0.797	0.361*	0.427*

Table 5: Fraction of NPs and Entities in **Left** that are not in **Right** & vice-versa. * indicates Kendall τ with $p < 0.05$.

vestigate whether the performance of KPA models reflects this subjectivity. That is, we test if match scores x correlate with the agreement between annotators. Intuitively, when annotators agree, an argument and key point should be considered to match more objectively and thus may be easier to score for a model. From the two datasets that have a per-sample agreement score, we measure the Pearson r correlation between the annotator agreement percentage (as obtained from data) and each approach's match score $M(d, k)$. Results are shown in Table 6.

	PVE		Perspectrum	
Approach	r	p	r	p
ChatGPT	0.030	0.687	0.039	0.469
Debater	0.163	0.029	-0.051	0.013
SMatch-base	0.097	0.195	0.093	0.215
SMatch-large	0.207	0.005	-0.03	0.123

Table 6: Pearson r correlation scores between predicted match scores and the annotator agreement per sample.

For all approaches, the correlations are negligible or weak at best (Schober et al., 2018). This shows that the predictions made by the models fail to identify which matches are interpreted differently among annotators. Hence, these models are not able to represent the diversity stemming from annotation accurately (Plank, 2022).

Data sources The KPG and KPM evaluations (Sections 5.1 and 5.2) indicate how the methods perform when applied to different datasets. The performance is dataset- and task-specific; no single approach performs both tasks best on any dataset.

We further investigate the data sources in the Perspectrum dataset, which was constructed using three distinct sources. Figure 2 shows the performance on each source separately. Although ArgKP and Perspectrum share a data source, we find no overlapping claims and little repetition in content between the two (App. A.1).

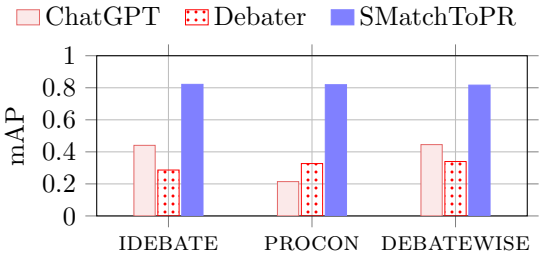

Figure 2: KPM performance for all approaches on the different data sources in Perspectrum.

The SMatchToPR and Debater approaches are not sensitive to data source shift, but ChatGPT performance differs depending on the source data used, dropping considerably for the *procon* source. We find two factors that influence why these arguments are harder to match: (1) *procon* contains about 10 times fewer claims than the other two sources, and (2) *procon*'s arguments are copied verbatim from various cited sources, leading to large stylistic and argumentative differences.

6 Conclusion

We perform a novel diversity exploration of different KPA approaches on three distinct datasets. By splitting KPA into two subtasks (KPG and KPM), we investigate each subtask, independently.

First, we find that an LLM-based approach works well for generating key points, but fails to match arguments to key points reliably. Conversely, smaller fine-tuned models are better at matching arguments to key points but struggle to find good key points consistently. Second, using a single training set yields poor generalization across datasets, showing that data source impacts a KPA approach's ability to generalize. Diversification of training data leads to promising results. Third, across all datasets, we see that existing methods for KPA are insensitive to long tail diversity, decreasing performance for key points supported by few arguments.

Finally, all models are insensitive to differences between individual annotators, disregarding subjective interpretations of arguments and key points.

We showed how multiple aspects of diversity, a core principle when interpreting opinions, are not evaluated using the standard set of metrics. Our analysis revealed interesting complementary strengths of the KPA approaches. Future efforts could focus on addressing diversity, either by mining for minority opinions directly (Waterschoot et al., 2022), or by identifying possibly subjective instances using socio-demographic information (Beck et al., 2023). Further, models can be enhanced with subjective understanding (Romberg, 2022; van der Meer et al., 2023), or work together with humans to jointly address some of the diversity issues (van der Meer et al., 2022; Argyle et al., 2023).

Ethics Statement

There are growing ethical concerns about NLP (broadly, AI) technology, especially, when the technology is used in sensitive applications. Argument summarization can be used in sensitive applications, e.g., to assist in public policy making. An ethical scrutiny of such methods is necessary before their societal application. Our work contributes toward such scrutiny. The outcome of our analysis shows how KPA methods fail to handle diversity. Potential technological improvements may lead to better results, but due diligence is required before applying such methods to real-world use cases.

We do not collect new data or involve human subjects in this work. Thus, we do not introduce any ethical considerations regarding data collection beyond those that affect the original datasets. A potential concern is that reproducing our results may involve using (possibly paid) services for running KPA. However, we aimed to make the analyses feasible with limited budget and resources.

Limitations

We identify five limitations of our work.

Diversity definition Our definition of diversity is specific to three dimensions, but there may be additional dimensions. For example, our unit of analysis is at the *argument* level. Diversity may also be analyzed for the opinion holders or those affected by decisions in policy-making contexts.

Novel key points Our evaluation of KPG and KPM employs existing key points. However, KPA methods may generate novel or unseen key points. Evaluating such novel key points is nontrivial and it may require experiments involving human subjects.

Resource limitations KPA approaches are resource intensive. We limited some approaches where (1) it would become too expensive to run KPA because of the complexity of the number of comparisons (e.g., Debater approach), or (2) the models do not support a big enough window to fit all arguments (e.g., ChatGPT context window is limited). While there are alternatives (e.g., GPT-4), they drastically increase the cost.

Dataset diversity The arguments in our data are in English, and limited to data gathered from online sources. Further, the users involved in collecting the datasets we employ may not be demographically representative of the global population. We conjecture that increasing the diversity of the data sources would make our conclusions stronger. However, publicly available datasets, especially non-English sources, for this task are scarce. We make our code and experimental data public to incentivize further research in this direction.

Data exposure We cannot verify whether the data from the test sets have been used when training the LLMs. This would make the model familiar with the vocabulary and have a more reliable estimation of the arguments' semantics. That likelihood is the smallest for PVE since it is the most recent dataset, gathered with new crowd workers.

Acknowledgements

This research was funded by the Netherlands Organisation for Scientific Research (NWO) through the Hybrid Intelligence Centre via the Zwaartekracht grant (024.004.022). We would like to thank the ARR reviewers for their feedback.

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.

Milad Alshomary, Jonas Rieskamp, and Henning Wachsmuth. 2022. Generating contrastive snippets

for argument search. In *Computational Models of Argument*, pages 21–31. IOS Press.

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Aspect-controllable opinion summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6578–6593.

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293.

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686, Brussels, Belgium. Association for Computational Linguistics.

Miguel Arana-Catania, Felix-Anselm Van Lier, Rob Procter, Nataliya Tkachenko, Yulan He, Arkaitz Zubiaga, and Maria Liakata. 2021. Citizen participation and machine learning for a better democracy. *Digital Government: Research and Practice*, 2(3):1–22.

Nikolay Archak, Anindya Ghose, and Panagiotis G Ipeirotis. 2007. Show me the money! deriving the pricing power of product features by mining consumer reviews. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 56–65.

Lisa P. Argyle, Christopher A. Bail, Ethan C. Busby, Joshua R. Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. 2023. Leveraging ai for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences*, 120(41):e2311627120.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020a. From arguments to key points: Towards automatic argument summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Roy Bar-Haim, Lilach Eden, Yoav Kantor, Roni Friedman, and Noam Slonim. 2021a. Every bite is an experience: Key point analysis of business reviews. 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 3376–3386.

Roy Bar-Haim, Yoav Kantor, Lilach Eden, Roni Friedman, Dan Lahav, and Noam Slonim. 2020b. Quantitative argument summarization and beyond: Cross-domain key point analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 39–49.

Roy Bar-Haim, Yoav Kantor, Elad Venezian, Yoav Katz, and Noam Slonim. 2021b. Project Debater APIs: Decomposing the AI grand challenge. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 267–274, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Eric PS Baumer, Mahmood Jasim, Ali Sarvghad, and Narges Mahyar. 2022. Of course it's political! a critical inquiry into underemphasized dimensions in civic text visualization. In *Computer Graphics Forum*, 3, pages 1–14. Wiley Online Library.

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

Surbhi Bhatia. 2020. A comparative study of opinion summarization techniques. *IEEE Transactions on Computational Social Systems*, 8(1):110–117.

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.

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.

Paul Cairney. 2016. *The politics of evidence-based policy making*. Springer.

Arie Cattan, Lilach Eden, Yoav Kantor, and Roy Bar-Haim. 2023. From key points to key point hierarchy: Structured and expressive opinion summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 912–928, Toronto, Canada. Association for Computational Linguistics.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle: Discovering diverse perspectives about claims. 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 542–557, Minneapolis, Minnesota. Association for Computational Linguistics.

Eric Chu and Peter Liu. 2019. Meansum: A neural model for unsupervised multi-document abstractive summarization. In *International Conference on Machine Learning*, pages 1223–1232. PMLR.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. In *International Conference on Learning Representations*.

Paul Clough, Robert Gaizauskas, Scott SL Piao, and Yorick Wilks. 2002. Measuring text reuse. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 152–159.

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

Charlie Egan, Advaith Siddharthan, and Adam Wyner. 2016. Summarising the points made in online political debates. In *Proceedings of the 3rd Workshop on Argument Mining, The 54th Annual Meeting of the Association for Computational Linguistics*, pages 134–143. Association for Computational Linguistics (ACL).

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert systems with applications*, 165:113679.

Roni Friedman-Melamed, Lena Dankin, Yufang Hou, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2021. Overview of the 2021 key point analysis shared task. In *Conference on Empirical Methods in Natural Language Processing*.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Shai Gretz, Assaf Toledo, Roni Friedman, Dan Lahav, Rose Weeks, Naor Bar-zeev, João Sedoc, Pooja Sangha, Yoav Katz, and Noam Slonim. 2023. Benchmark data and evaluation framework for intent discovery around covid-19 vaccine hesitancy. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1328–1340.

Max Grusky. 2023. Rogue scores. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1914–1934.

Ali Gürkan, Luca Iandoli, Mark Klein, and Giuseppe Zollo. 2010. Mediating debate through on-line large-scale argumentation: Evidence from the field. *Information Sciences*, 180(19):3686–3702.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

J. Edward Hu, Abhinav Singh, Nils Holzenberger, Matt Post, and Benjamin Van Durme. 2019. Large-scale, diverse, paraphrastic bitexts via sampling and clustering. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 44–54, Hong Kong, China. Association for Computational Linguistics.

Nannan Huang, Lin Tian, Haytham Fayek, and Xiuzhen Zhang. 2023. Examining bias in opinion summarisation through the perspective of opinion diversity. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 149–161, Toronto, Canada. Association for Computational Linguistics.

IBM Research. 2023. The Project Debater Service API. `https://developer.ibm.com/apis/catalog/debater--project-debater-service-api/Introduction`. Accessed: October 2023.

David Inouye and Jugal K Kalita. 2011. Comparing twitter summarization algorithms for multiple post summaries. In *2011 IEEE Third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 298–306. IEEE.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

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

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

Hao Li, Viktor Schlegel, Riza Batista-Navarro, and Goran Nenadic. 2023. Do you hear the people sing? key point analysis via iterative clustering and abstractive summarisation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14064–14080, Toronto, Canada. Association for Computational Linguistics.

Enrico Liscio, Roger Lera-Leri, Filippo Bistaffa, Roel I.J. Dobbe, Catholijn M. Jonker, Maite Lopez-Sanchez, Juan A Rodriguez-Aguilar, and Pradeep K Murukannaiah. 2023. Value inference in sociotechnical systems: Blue sky ideas track. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems, AAMAS*, volume 23, pages 1–7.

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*.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Jane Mansbridge, James Bohman, Simone Chambers, Thomas Christiano, Archon Fung, John Parkinson, Dennis F Thompson, and Mark E Warren. 2012. A systemic approach to deliberative democracy. *Deliberative systems: Deliberative democracy at the large scale*, pages 1–26.

Duncan McLaren, Karen A Parkhill, Adam Corner, Naomi E Vaughan, and Nicholas F Pidgeon. 2016. Public conceptions of justice in climate engineering: Evidence from secondary analysis of public deliberation. *Global Environmental Change*, 41:64–73.

Amita Misra, Pranav Anand, Jean E Fox Tree, and Marilyn Walker. 2015. Using summarization to discover argument facets in online idealogical dialog. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 430–440.

Niek Mouter, Jose Ignacio Hernandez, and Anatol Valerian Itten. 2021. Public participation in crisis policymaking. how 30,000 dutch citizens advised their government on relaxing covid-19 lockdown measures. *PLoS One*, 16(5):e0250614.

Eni Mustafaraj, Samantha Finn, Carolyn Whitlock, and Panagiotis T Metaxas. 2011. Vocal minority versus silent majority: Discovering the opionions of the long tail. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 103–110. IEEE.

Yoon-Eui Nahm. 2013. A novel approach to prioritize customer requirements in qfd based on customer satisfaction function for customer-oriented product design. *Journal of Mechanical Science and Technology*, 27:3765–3777.

Ani Nenkova, Kathleen McKeown, et al. 2011. Automatic summarization. *Foundations and Trends® in Information Retrieval*, 5(2–3):103–233.

OpenAI. 2023. The OpenAI Python library. `https://github.com/openai/openai-python`. Accessed: October 2023.

Hoang Phan, Long Nguyen, and Khanh Doan. 2021. Matching the statements: A simple and accurate model for key point analysis. In *Proceedings of the 8th Workshop on Argument Mining*, pages 165–174.

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.

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.

Ahnaf Mozib Samin, Behrooz Nikandish, and Jingyan Chen. 2022. Arguments to key points mapping with prompt-based learning. *ICNLSP 2022*, page 303.

Patrick Schober, Christa Boer, and Lothar A Schwarte. 2018. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5):1763–1768.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

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.

Tejpalsingh Siledar, Jigar Makwana, and Pushpak Bhattacharyya. 2023. Aspect-sentiment-based opinion summarization using multiple information sources. In *Proceedings of the 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD)*, pages 55–61.

Julius Sim and Chris C Wright. 2005. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy*, 85(3):257–268.

Yoshihiko Suhara, Xiaolan Wang, Stefanos Angelidis, and Wang-Chiew Tan. 2020. Opiniondigest: A simple framework for opinion summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5789–5798.

James Surowiecki. 2005. *The wisdom of crowds*. Anchor.

Shahbaz Syed, Khalid Al Khatib, Milad Alshomary, Henning Wachsmuth, and Martin Potthast. 2021. Generating informative conclusions for argumentative texts. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3482–3493.

Michiel van der Meer, Enrico Liscio, Catholijn M Jonker, Aske Plaat, Piek Vossen, and Pradeep K Murukannaiah. 2022. HyEnA: A hybrid method for extracting arguments from opinions. In *HHAI2022: Augmenting Human Intellect*, pages 17–31. IOS Press.

Michiel van der Meer, Piek Vossen, Catholijn Jonker, and Pradeep Murukannaiah. 2023. Do differences in values influence disagreements in online discussions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15986–16008, Singapore. Association for Computational Linguistics.

Michiel van der Meer, Piek Vossen, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2024. An empirical analysis of diversity in argument summarization: Supplementary material.

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.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 47–57, San Diego, California. Association for Computational Linguistics.

Cedric Waterschoot, Ernst van den Hemel, and Antal van den Bosch. 2022. Detecting minority arguments for mutual understanding: A moderation tool for the online climate change debate. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6715–6725.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Detailed Experimental Setup

We describe our experimental setup, starting with the data we use for conducting our analysis. We follow with a detailed description of each approach and finally present a description of the metrics used.

A.1 Data

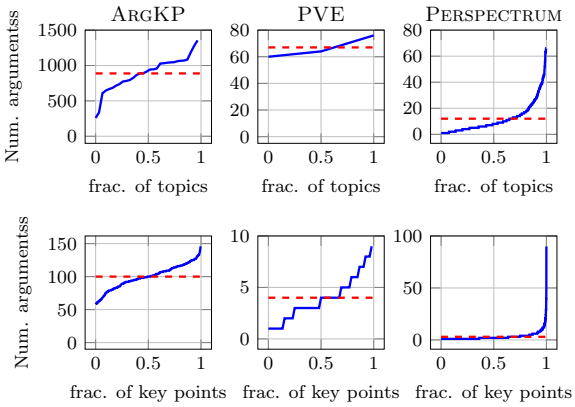

Figure 3: Number of arguments matched per claim (upper row) and key point (bottom row), sorted by frequency. The red dashed line shows the average number of arguments.

We provide some quantitative statistics on the three datasets used in our work in Table 7. In addition, we show some qualitative examples of the content in our datasets in Table 9. Since Perspectrum and ArgKP listed the same debate platforms as sources, we investigate the overlap between the claims and arguments between pairs of datasets. In terms of claims, there is no direct overlap between any two datasets. To rule out that the same arguments were scraped from the debate platforms, we also measure n-gram overlap (Clough et al., 2002). We show the overlap in unigrams, bigrams, and trigrams in Table 8. The overlap scores report the ratio of n-grams from one dataset that is found in the other.

For PVE, since the key point analysis was performed using a mixture of crowd and AI techniques, we take only the correctly matched key point–motivation pairs. That is, we take only those pairs that were deemed matching according to the final evaluation performed.

A.2 Per-approach Specifics

See Table 10 for the language models used in each approach. We further outline any details depending on the approach used.

Dataset	Num. arguments	Num. Key Points	Num. claims	Avg. arguments per claim	Avg. arguments per KP
ARGKP	10717	277	31	245	20
PVE	269	185	3	67	4
PERSPECTRUM	10927	3804	905	12	3

Table 7: Quantitative statistics of the datasets used in the experiments.

		Target		
		ARGKP	PVE	PERSPECTRUM
Source	ARGKP	–	0.40/0.08/0.01	0.70/0.21/0.14
	PVE	0.41/0.16/0.06	–	0.66/0.24/0.10
	PERSPECTRUM	0.17/0.04/0.02	0.22/0.03/0.01	–

Table 8: Maximum uni-, bi-, and trigram overlap between datasets.

Debater The Debater API allows multiple parameters when running the KPA analysis. We manually tuned the parameters separately for KPG and KPM. For both tasks, we started with the most permissive configuration to optimize for recall first, and gradually made parameters more strict to improve precision without lowering recall scores. Once recall scores started dropping, we fixed the parameters. The final configuration is shown in Table 11.

ChatGPT We strive to make our results as reproducible as possible, but due to the nature of the OpenAI API results may be specific to model availability. We conducted the experiments between July and August 2023, using the `gpt-3.5-turbo` and `gpt-3.5-turbo-16k` models. We provide a template for the prompts below, in Prompts 1, 2, and 3. Open-book ChatGPT for KPG uses up to $B_{KPG} = 600, 100, 100$ for ARGKP, PVE, PERSPECTRUM respectively. ChatGPT uses a batch size of $B_{KPM} = 10$ when making match predictions for KPM.

Interpreting the responses was done by prompting the model to output valid JSON, and writing a script that parses the generated response. Invalid JSON responses are considered errors on the model's side, resulting in an empty string for KPG and a 'no-match' label for KPM. In order to cut down on costs, we subsampled the test set for PERSPECTRUM, taking a random 15% of the claims in order to drive down the costs further.

Prompt 1: ChatGPT closed book, KPG prompt

Give me a JSON object of key arguments for and against the claim: **{claim}**. Make sure the reasons start with addressing the main point. Indicate per reason whether it supports (pro) or opposes (con) the claim. Rank all reasons from most to least popular. Make sure you generate a valid JSON object. The object should contain a list of dicts containing fields: 'reason' (str), 'popularity' (int), and 'stance' (str).

Prompt 2: ChatGPT open book, KPG prompt

Extract key arguments for and against the claim: **{claim}**. You need to extract the key arguments from the comments listed here: **{up to B_{KPG} arguments}** Give me a JSON object of key arguments for and against the claim. Make sure the reasons start with addressing the main point. Indicate per reason whether it supports (pro) or opposes (con) the claim. Rank all reasons from most to least popular. Make sure you generate a valid JSON object. The object should contain a list of dicts containing fields: 'reason' (str), 'popularity' (int), and 'stance' (str).

Prompt 3: ChatGPT open book, KPM prompt

For the claim of **{claim}**, indicate for each of the following argument/key point pairs whether the argument matches the key point. Return a JSON object with just a "match" boolean per argument/key point pair.

ID: **{pair id}** Argument: **{argument}** Key point: **{key point}** (*up to B_{KPM} times*) ...

SMatchToPR We preprocess the PERSPECTRUM dataset analogously to the ARGKP dataset. We train the SMatchToPR model using contrastive loss for 10 epochs and a batch size of 32. The training has a warmup phase of the first 10% of data.

Dataset	Claim	Key Point	Argument
ArgKP	We should subsidize journalism	Journalism is important to information-spreading/accountability.	Journalism should be subsidized because democracy can only function if the electorate is well informed.
PVE	Young people may come together in small groups	Young people are at low risk of getting infected with COVID-19 and therefore can benefit from gathering together with limited risk and potential profit.	Risks of contamination or transfer have so far been found to be much smaller.
Persp.	The threat of Climate Change is exaggerated	Overwhelming scientific consensus says human activity is primarily responsible for global climate change.	The biggest collection of specialist scientists in the world says that the world's climate is changing as a result of human activity. The scientific community almost unanimously agrees that man-caused global warming is a severe threat, and the evidence is stacking.

Table 9: Qualtitative examples of claims, key points and arguments across our dataset.

Name	Model
ChatGPT	`gpt-3.5-turbo-16k`
ChatGPT (closed book)	`gpt-3.5-turbo`
Debater	*closed-source*
SMatchToPR (base)	`RoBERTa-base`
SMatchToPR (large)	`RoBERTa-large`

Table 10: Models used for each KPA approach.

Parameter	Value
KPG	
mapping_policy	*LOOSE*
kp_granularity	*FINE*
kp_relative_aq_threshold	0.5
kp_min_len	0
kp_max_len	100
kp_min_kp_quality	0.5
KPM	
min_matches_per_kp	0
mapping_policy	*LOOSE*

Table 11: API Configuration for Debater approach.

The base and large variants use the same parameters. See Table 12 for the hyperparameters when executing KPG and KPM. The computing infrastructure used contained two RTX3090 Ti GPUs. Training the RoBERTa large variant takes around 30 minutes.

A.3 Evaluation metrics

KPG Since we use ROUGE scores for evaluation, to make our results reproducible we provide further details on the configuration of the ROUGE metrics (Grusky, 2023). Our evaluation uses the `sacrerouge` package that wraps the original ROUGE implementation[2]. The full evaluation parameters can be seen in Table 13.

Furthermore, we use two learned metrics (BLEURT and BARTScore) to report the semantic similarity of generated key points and reference key points. For BLEURT, we use the publicly available `BLEURT-20` model, which is a RemBERT (Chung et al., 2020) model trained on an augmented version of the WMT shared task data (Ma et al., 2019). BARTScore uses a BART model trained on ParaBank2 (Hu et al., 2019).

B Additional results

We present two additional results: we provide fine-grained ROUGE results for KPG, and provide examples of key points generated by ChatGPT.

[2] https://github.com/danieldeutsch/sacrerouge

Parameter	Value
PR d	0.2
PR min quality score	0.8
PR min match score	0.8
PR min length	5
PR max length	20
filter min match score	0.5
filter min result length	5
filter timeout	1000

Table 12: Hyperparameters for SMatchToPR approach.

Parameter	Value
Porter Stemmer	*yes*
Confidence Interval	95
Bootstrap samples	1000
α	0.5
Counting unit	*sentence*

Table 13: Configuration parameters for the ROUGE evaluation of KPG.

B.1 Detailed ROUGE scores for Key Point Generation

Earlier, we provided aggregated F_1 scores for the KPG evaluation. Here, we also show Precision and Recall scores in Table 14. We see that the models that perform best in terms of F_1 score are consistently scoring well in terms of precision and recall across all datasets. For instance, open-book ChatGPT performs best on ArgKP in terms of F_1 (see Table 3), achieving consistently high precision and recall scores. Other approaches may score higher on individual metrics (e.g. SMatchToPR large scores higher in terms of ROUGE-1 recall), but this pattern is not consistent across all metric types.

B.2 Additional BERTScores for Key Point Generation

Next to BLEURT and BARTScore, we report BERTScore (Zhang et al., 2020) for the approaches in the KPG evaluation, to examine the relation between the various learned metrics.

B.3 Long-tail experiment for KPG

We perform the long-tail analysis for Key Point Generation, adopting the same cutoff parameter f from the KPM analysis. Figure 4 shows the results when including a fraction of key points f, starting from the least frequent (i.e. the key points with the lowest amount of arguments matched to them). The figure shows that for a low fraction of data, all approaches perform considerably worse. Note that due to the evaluation setup in Li et al. (2023), scores may be lower due to a smaller pool of key points. Since we report averages of the maximum scoring match between any given generated and reference key points, this smaller pool may lead to overall lower scores. We still report these results to show the impact of making the evaluation set smaller, next to focusing on infrequent opinions.

B.4 ChatGPT generated key points for PVE

See Table 16. A cursory search for the content of the open-book key points shows the key points are directly taken from arguments in PVE. While ChatGPT performs conditioned language generation, it behaves like extractive summarization when using the open-book approach for the arguments in PVE. This leads to potentially incomplete or subjective key points. For the closed-book approach, we observe that ChatGPT generates independent and objective key points.

Data	Approach	Precision			Recall		
		R-1	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
ARGKP	ChatGPT	29.1	**10.6**	25.6	45.2	**16.1**	41.2
	ChatGPT (closed book)	**30.8**	6.8	**26.9**	32.0	8.6	27.3
	Debater	25.3	5.5	23.1	28.2	5.3	23.4
	SMatchToPR (base)	24.5	9.3	23.2	44.5	11.2	41.5
	SMatchToPR (large)	22.0	6.4	19.4	**53.0**	13.0	**47.5**
PVE	ChatGPT	25.1	6.4	21.1	19.1	3.9	15.8
	ChatGPT (closed book)	30.1	**9.8**	22.6	**26.4**	**8.1**	**21.6**
	Debater	**33.3**	0.0	**33.3**	13.3	7.1	13.3
	SMatchToPR (base)	28.8	5.6	22.6	18.0	2.9	14.4
	SMatchToPR (large)	27.8	5.6	22.6	18.0	2.9	14.4
PERSPECTRUM	ChatGPT	17.5	4.7	14.8	**35.0**	**10.2**	**30.5**
	ChatGPT (closed book)	14.8	3.1	12.8	25.4	6.3	22.7
	Debater	8.6	0.4	7.6	25.5	6.3	22.7
	SMatchToPR (base)	18.8	5.5	15.9	32.0	9.2	27.8
	SMatchToPR (large)	**19.0**	**5.7**	**16.1**	32.3	9.8	28.3

Table 14: ROUGE Precision and Recall scores for the Key Point Generation task.

Data	Approach	BERTScore		
		Precision	**Recall**	$\mathbf{F_1}$
ARGKP	ChatGPT	0.412	0.470	0.422
	ChatGPT (closed book)	0.322	0.336	0.324
	Debater	0.406	0.367	0.379
	SMatchToPR (base)	0.362	0.463	0.394
	SMatchToPR (large)	0.361	0.482	0.402
PVE	ChatGPT	0.184	0.157	0.153
	ChatGPT (closed book)	0.386	0.280	0.324
	Debater	0.523	0.146	0.301
	SMatchToPR (base)	0.339	0.210	0.257
	SMatchToPR (large)	0.339	0.210	0.257
PERSPECTRUM	ChatGPT	0.208	0.308	0.252
	ChatGPT (closed book)	0.244	0.274	0.243
	Debater	0.228	0.274	0.246
	SMatchToPR (base)	0.231	0.297	0.258
	SMatchToPR (large)	0.235	0.296	0.260

Table 15: BERTScore Precision, Recall, and F_1 scores for the Key Point Generation task.

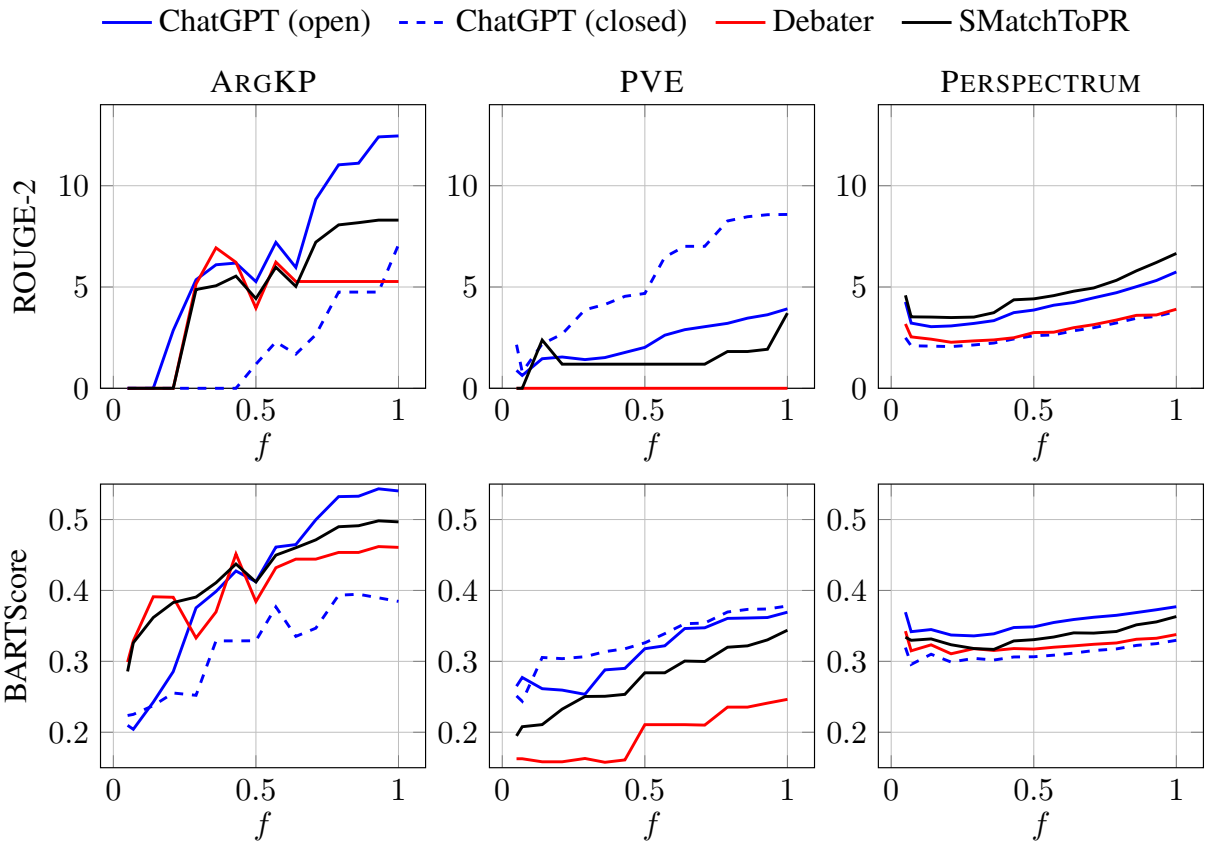

Figure 4: KPG performance when limiting data usage to a fraction f, starting with the long tail first.

Claim	**Stance**	**KP (open-book)**	**KP (closed-book)**
All restrictions are lifted for persons who are immune	con	The coronavirus is an assassin, let's really learn more about this first	There may still be unknown long-term effects of the virus, even in those who have recovered.
Re-open hospitality and entertainment industry	pro	Economy needs to start running again	Reopening the hospitality and entertainment industry will help stimulate the economy and create job opportunities.
Young people may come together in small groups	con	The spread will then come back in all its intensity.	Small group gatherings may pose a risk of spreading contagious diseases.

Table 16: Examples of generated key points from the open-book and closed-book ChatGPT approach.