

ARQUSUMM: Argument-aware Quantitative Summarization of Online Conversations

An Quang Tang, Xiuzhen Zhang ^{*}, Minh Ngoc Dinh, Zhuang Li

RMIT University, Australia

s3695273@rmit.edu.vn, xiuzhen.zhang@rmit.edu.au,
minh.dinh4@rmit.edu.vn, zhuang.li@rmit.edu.au

Abstract

Online conversations have become more prevalent on public discussion platforms (e.g. Reddit). With growing controversial topics, it is desirable to summarize not only diverse arguments, but also their rationale and justification. Early studies on text summarization focus on capturing general salient information in source documents, overlooking the argumentative nature of online conversations. Recent research on conversation summarization although considers the argumentative relationship among sentences, fail to explicate deeper argument structure within sentences for summarization. In this paper, we propose a novel task of argument-aware quantitative summarization to reveal the claim-reason structure of arguments in conversations, with quantities measuring argument strength. We further propose ARQUSUMM, a novel framework to address the task. To reveal the underlying argument structure within sentences, ARQUSUMM leverages LLM few-shot learning grounded in the argumentation theory to identify propositions within sentences and their claim-reason relationships. For quantitative summarization, ARQUSUMM employs argument structure-aware clustering algorithms to aggregate arguments and quantify their support. Experiments show that ARQUSUMM outperforms existing conversation and quantitative summarization models and generate summaries representing argument structures that are more helpful to users, of high textual quality and quantification accuracy.

Code — <https://github.com/antangrocket1312/ArQuSumm>

Introduction

The proliferation of user participation of online conversations platforms such as online discussion forums (e.g., Reddit¹) (Völske et al. 2017; Zhang et al. 2019), community question answering sites and online news discussion forums has resulted in large volumes of online conversation data. Summarization of online conversations has become an important text summarization task (Fabbri et al. 2021). Different from well structured documents such as news articles or scientific papers, online conversations, with many users frequently include hundreds of arguments, can spread

across the various threads of online conversations (Syed et al. 2023). In fact, users not only engage in discussions to express their viewpoints, but also debate, justify, and challenge others’ viewpoints. User comments are therefore often argumentative, with inferential relations among propositions expressing not only what people believe, but also why.

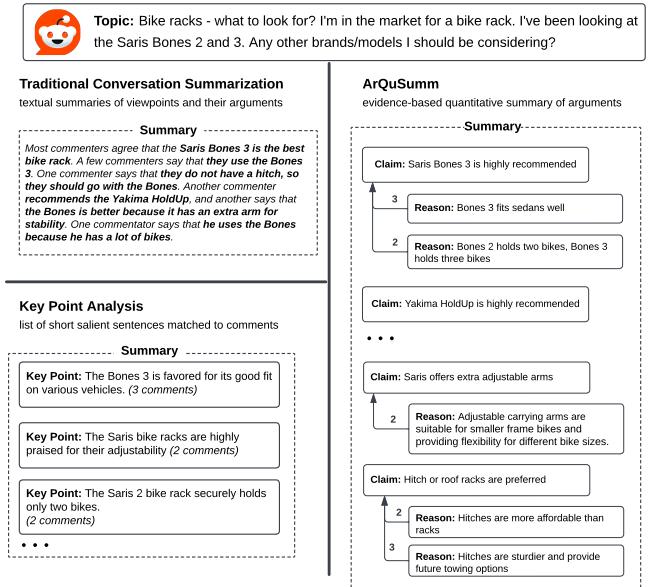


Figure 1: Comparison of ARQUSUMM, and existing conversation and quantitative summarization methods

Conventional multi-document textual summarization aims to output the most salient parts of multiple related documents in a concise and readable form. Early research on conversation summarization utilized document summarization models to summarize conversations (Liu and Lapata 2019; Lewis et al. 2020; Zhang et al. 2020a), but success is limited because conversational text contains main points scattering across multiple utterances and between numerous writers (Gliwa et al. 2019). Recent studies then adopted argumentation theories (Barker and Gaizauskas 2016) and frameworks (Barker et al. 2016) to model arguments and viewpoints presented in the conversation for summarization (Lenz et al. 2020; Chen and Yang 2021; Fabbri et al.

^{*}Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹<https://www.reddit.com/>

2021). Still, existing studies fail to capture and comprehensively present the rationale and justification for arguments in their plain textual summary. Further, arguments are modelled at the sentence level where sentences are the argument elements, often resulting in redundancy and incoherence in argumentative logic across sentences. Importantly, the plain summary text lacks the ability to explicitly represent the argument structure and explain the reasons for claims.

Recently the quantitative view was introduced into textual summaries to capture and numerically quantify diverse viewpoints in reviews and debates, in a task known as Key Point Analysis (KPA) (Bar-Haim et al. 2020a,b, 2021; Tang, Zhang, and Dinh 2024; Tang et al. 2024). KPA summarizes user comments (sentences) into concise sentences called key points (KPs), which are claims expressing user viewpoints, and quantify their prevalence. Nevertheless, KPA only generate a summary as a list of claims without capturing the argumentative logic between them, leaving to users to assess the reliability of these KPs on their own. Moreover, the quantification of prevalence for KPs is mostly via matching KPs with users comments based on only semantic similarity in texts, which can be inaccurate.

In this paper, we propose a novel framework ARQSUMM for argument-aware quantitative summarization of conversations. Different from prior studies, the summary includes *claim-reason argument structures* explaining arguments in the summary. Figure 1 shows an example. When discussing “which bike rack models to look for”, ARQSUMM presents every argument directly as a tree structure; the root represents the claim from user comments (e.g. *should use Saris Bone 3*), and each leaf represents a reason for the claim (e.g., *Bone 3 fits sedan well*), along with its support quantity (number of comments).

Specifically, to explicate the argument structure in user comments, the ARQSUMM framework first leverages LLM in-context learning grounded in argumentation theory to identify propositions as argument elements within sentences. It then employs LLM in-context learning to predict the entailment relationship and identify claim and reason for the structure of arguments. Importantly, we propose an argument structure-aware clustering algorithms to aggregate fine-grained propositions into distinct high-level claims and quantify their supporting reasons effectively.

Our main contributions are:

- We introduce the novel task of argument-aware quantitative summarization of online conversations, where, unlike conventional plain textual summaries, argument structures are explicitly represented. Experiments show that our new form of summary offers 7 times more comprehensive and useful presentation, and our proposed framework outperforms baselines with up to 3.71 times improvement in textual similarity with ground-truth arguments and up to 0.421 improvement in F1 matching over current quantitative summarization framework (Tang et al. 2024).
- Different from existing studies we leverage LLM in-context learning grounded in argumentation theory to identify text spans of argument elements in sentences.

- Leveraging LLM in-context learning to identify claim-reason relations for arguments, we design novel argument structure-aware clustering algorithms to form and quantify arguments and generate the structured summary.

Related Work

Textual Summarization of Conversations Existing studies focus on textual summarization of conversations, utilizing the conversational structure. Barker et al. (2016) proposed conversation overview summary to capture the key contents of reader comment conversations for news articles. Misra et al. (2017) use summarization to discover central propositions in online debates. Barker and Gaizauskas (2016) identify three key components of conversational dialogues to manually construct an argument graph for the whole conversation. Building on this theoretical framework for argumentation, Fabbri et al. (2021) applied entailment relations to automatically construct argument graphs for conversations and generate textual summaries. Note that these existing studies model argument relations at the sentence level, which is a loose assumption according to the argumentation theory (Toulmin 1958).

Quantitative Summarization of Key Points Key Point Analysis (KPA) was recently proposed to summarize key points (KPs) quantify them for reviews and debates (Bar-Haim et al. 2020a,b). To address that a flat list of KPs often expresses related ideas at varying levels of granularity, Cattan et al. (2023) proposes to summarize key points into a hierarchy. But their pipeline approach of first generating and then summarizing KPs can give rise to cascading of errors, leading to poor KP hierarchies.

Argument Mining Recent argument mining studies focus on identifying argumentative units and structures based on argumentation theories (Stab and Gurevych 2014). Gupta, Zuckerman, and O’Connor (2024) proposes a quantitative argumentation framework (QAF) that harnesses LLMs and Toulmin theory (Toulmin 1958) to explicate and cluster arguments from comments into a hypergraph. However, these graphs lack conciseness and is cluttered with multiple layers of arguments, with some being possibly overlapping due to surface-level argument clustering approach. Still, existing studies can only produce an argument graph to visualize the argument flow rather than a concise and readable summary.

Task Formulation

Let $\mathcal{D} = \{d_i\}_{i=1}^{|\mathcal{D}|}$ denote a collection of input comments in an online discussion thread. From each comment d_i , we extract a set of argument propositions, namely a claim c and its associated reasons $\{r_j\}$. This results in a set of extracted propositions for \mathcal{D} , denoted as $\mathcal{P} = \{(c_i, \mathcal{R}_i)\}_{i=1}^{|\mathcal{P}|}$, where each c_i is a claim and \mathcal{R}_i is a set of reasons justifying it. In the the Argument-aware Quantitative Summarization (ARQSUMM) task, we aim generate a structured summary \mathcal{S} , which consists of a set of *claims* $\hat{\mathcal{C}} = \{\hat{c}_k\}_{k=1}^K$, each supported by a set of aggregated *reasons* $\hat{\mathcal{R}}_k = \{\hat{r}_{k1}, \hat{r}_{k2}, \dots\}$.

Formally, we define our summary as:

$$\mathcal{S} = \left\{ \left(\hat{c}_1, \hat{\mathcal{R}}_1 \right), \left(\hat{c}_2, \hat{\mathcal{R}}_2 \right), \dots, \left(\hat{c}_k, \hat{\mathcal{R}}_k \right) \right\}_{k=1}^K$$

where each \hat{c}_k represents a semantically unified cluster of input claims from \mathcal{P} , and each $\hat{\mathcal{R}}_k$ corresponds to a distinct subgroup of reasons drawn from the reason sets \mathcal{R}_i originally linked to the clustered claims. For simplicity, hereafter we refer to \hat{c}_k and $\hat{\mathcal{R}}_k$ produced in the final structured summary as *claim* and *reason*. For example, a final summary entry may take the form:

Claim: *Remote work improves productivity*

→ **Reason:** *Less time is wasted on commuting* (22 instances)

→ **Reason:** *Home environment allows for fewer distractions* (14 instances)

The ARQUSUMM framework

Figure 2 illustrates the overall pipeline of our proposed ARQUSUMM framework. Given a set of comments from an online discussion, ARQUSUMM performs argument-aware quantitative summarization by generating a structured summary of claims and their supporting reasons, each annotated with prevalence. The framework consists of three stages: *i*) *Argument Proposition Extraction*, *ii*) *Claim-Reason Clustering*, and *iii*) *Structured Summary Generation*.

Argument Proposition Extraction

Unlike previous summarization studies, we ground our summarization process on claims and reasons, i.e., text phrases, implied by the comments. The process is more effective and faithful than summarizing and quantifying on comment sentences because it can (1) bypass noise in the original sentence, (2) recover the original entity name in case of cross-sentence references (e.g., John instead of he). We specifically harnesses the LLM’s extensive knowledge on argumentation theory, by zero-shot prompting it to extract possible argument components inside a comment. Based on the experiment of various theories (Toulmin 1958; Walton 1996; Freeman 1991) for argument explication, we decided to make use of Toulmin (Toulmin 1958) due to LLM’s exceptional interpretation and capability to extract arguments following this theory (Gupta, Zuckerman, and O’Connor 2024). In particular, we prompt an LLM with references to Toulmin’s theory (e.g., ‘According to Toulmin model,’) (Gupta, Zuckerman, and O’Connor 2024), which elicits a response that correctly bases on Toulmin’s theory to extract values from the input comment corresponding to the ‘claim’ and ‘reason’ from the theory. The Toulmin theory decomposes the arguments within a comment into three components, namely: **claim** (assertion or viewpoint made by the author for general acceptance), **reason** (proposition provided by the author to convince the audience to accept the claim), and **warrant** (the author’s world knowledge explain why the claim follow from the provided reason). Note that we omit the warrant layer, generated by LLM’s parametric knowledge, to preserve original user reasoning. Note also that arguments in a comment (with multiple sentences) can

carry multiple claims, where each claim could be supported by multiple reasons.

Claim-Reason Clustering

Previous approaches (Gupta, Zuckerman, and O’Connor 2024) clustered argument propositions based on their embeddings irrespective of their role in the argument structure, which impairs the alignment between claims and their associated reasons and fails to explicitly preserve the inferential relationship between them. A key innovation of ARQUSUMM lies in disentangling and separately clustering claims and reasons to better reflect the underlying argument structure. We therefore propose a two-stage hierarchical clustering process grounded in argumentation theory and entailment-based reasoning.

Claim Clustering At the highest level of social discussion, identifying distinct claims to represent diverse viewpoints is central to argument summarization. Prior approaches often rely on semantic similarity of claims and reasons—typically using cosine distance in embedding space—to induce viewpoint clusters (Gupta, Zuckerman, and O’Connor 2024). However, such surface-level similarity overlooks critical differences in attitude, stance, and implied judgment that distinguish opposing viewpoints. In this work, we argue that clustering based solely on claim embeddings is insufficient for capturing meaningful argumentative distinctions. Instead, we propose leveraging entailment relationships between argument propositions as a proxy for viewpoint alignment.

Associated Reasons Distributional Entailment Clustering claims solely based on claim’s mutual entailment score might be ineffective and unreliable, while highly relevant claims might still have associated reasons unsupportive to each other. To mitigate this risk, we propose incorporating information from each claim’s associated reasons as a regularizing signal in the clustering process. Our method draws inspiration from the distributional inclusion hypothesis (Geffet and Dagan 2005), which suggests that the context surrounding an entailing word w_1 is naturally expected to occur also with the entailed word w_2 (Geffet and Dagan 2004). We adapt this hypothesis to the argumentative domain with the intuition that if claim c_m supports claim c_n , it is likely that a reason $r_j \in \mathcal{R}_m$ that supports c_m will also support c_n . We formalize this as the **Associated Reasons Distributional Entailment (ARDE)** score. Given a claim c_m and its associated reasons \mathcal{R}_m , we compute the proportion of reasons that also support a neighbouring claim c_n as:

$$\text{ARDE}(c_m \Rightarrow c_n) = \frac{|\{r_j \in \mathcal{R}_m : r_j \text{ supports } c_n\}|}{|\mathcal{R}_m|}$$

where a reason r_j of claim c_m is determined to support adjacent claim c_n if it exceeds a threshold t .

To cluster claims into distinctive groups, we construct a *fully connected graph* by computing two types of pairwise scores between all claims $c_i \in \mathcal{P}$: (1) the *entailment score*, and (2) the ARDE score between claim pairs in both directions. These scores are then combined to build a undirected

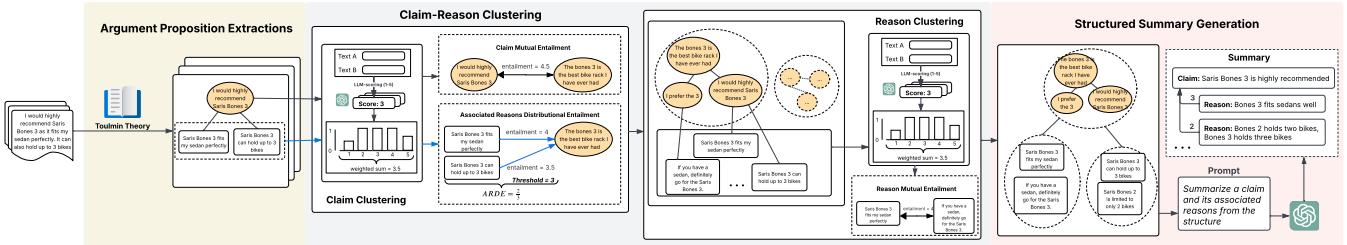


Figure 2: The ARQUSUMM framework.

graph G_{claim} , where nodes represent claims and edges capture mutual support relationships. Specifically, for each candidate pair (c_m, c_n) , we compute a final alignment score $s(m, n)$ as the average of the *bidirectional entailment scores* and *bidirectional ARDE scores*².

$$s(m, n) = \frac{1}{2} (\text{Entail}_{\text{bi}}(c_m, c_n) + \text{ARDE}_{\text{bi}}(c_m, c_n))$$

An edge $e(m, n)$ is added to G_{claim} if $s(m, n) > \tau$, where τ is a threshold controlling clustering granularity. Claims within the same *strongly connected component* of G are contracted into the same cluster, as they are mutually supportive in both semantic content and reasoning structure.

Aggregation and Clustering of Reasons After forming clusters of semantically equivalent claims, we aggregate their associated reasons to form a set of supporting evidence for each claim group. Specifically, for a claim cluster \hat{c}_k , we define its aggregated reason set as: $\hat{\mathcal{R}}_k = \bigcup_{c_i \in \text{cluster}_k} \mathcal{R}_i$. Next, we perform *reason clustering* within each $\hat{\mathcal{R}}_k$ to identify distinct subgroups of justifications. Each resulting subgroup represents a coherent reasoning pattern frequently cited in support of the aggregated claim. To determine whether two reasons $r_u, r_v \in \hat{\mathcal{R}}_k$ belong to the same cluster, we build a bipartite entailment graph to compute their *pairwise entailment scores* in both directions. Based on these scores, we construct an undirected graph $G = (\hat{\mathcal{R}}_k, E)$, where each node corresponds to a reason within $\hat{\mathcal{R}}_k$, and an undirected edge $\{r_u, r_v\} \in E$ is added if and only if average entailment score from both direction exceed a threshold τ . We then contract each strongly connected component of G to be a *reason cluster*, which represents a distinct subgroup of justification commonly used to support claim \hat{c}_k from online discussion.

LLM-based Entailment Scoring Function Recent studies suggest using LLMs as reference-free metrics for NLG evaluation. Inspired by the G-Eval LLM-based evaluator (Liu et al. 2023) for NLG output, we adapt this evaluator to the Natural Language Inference (NLI) task for scoring the entailment relationship between argument propositions extracted from social comments. Specifically, we utilize the probabilities of output tokens from LLMs to normalize the

²we average the pairwise score from each direction to obtain the bidirectional score

scores and take their weighted summation as the final results. Formally, given a set of scores (like from 1 to 5) pre-defined in the prompt $S = \{s_1, s_2, \dots, s_n\}$, the probability of each score $p(s_i)$ is calculated by the LLM, and the final entailment score is:

$$\text{score} = \sum_{i=1}^n p(s_i) \times s_i \quad (1)$$

The intuition is prompting the LLM to score the two given texts multiple times and then take the average of all runs to achieve fine-grained, continuous scores that better reflect the entailment relationship between argument propositions. Empirical validation shows that our LLM-based entailment scoring (using GPT-4.1) obtain stronger alignment with human judgement than scores produced by the state-of-the-art RoBERTa-large model³ ($r = 0.556$, Accuracy = 0.715)⁴.

Structured Summary Generation

Given the final set of claim clusters $\{\hat{c}_k\}_{k=1}^K$ and their corresponding reason clusters $\{\hat{\mathcal{R}}_k\}_{k=1}^K$, we generate the bullet-like structured summary S where each bullet is a tree structure, with its stem node representing \hat{c}_k and the connected leaf nodes representing $\hat{\mathcal{R}}_k$. Note that due to structural complexity and to ensure highly-coherent content, we prompt an LLM to summarize each claim cluster and its corresponding reason clusters for each argument in the discussion per generation, before aggregating output across all arguments to achieve a final claim-reason structure summary. Note also that to minimize ambiguity and hallucination, we explicitly structure the input of \hat{c}_k and $\hat{\mathcal{R}}_k$ as a JSON object. Importantly, we assign each claim and reason group with a unique identifier at input, and enforce the LM to include reference these IDs at output to ensure accurate alignment between summarized claims and reasons and their source clusters.

Prompt Engineering Following best practices from OpenAI’s prompt engineering guidelines⁵, we design the prompt with three key components: (1) a brief task description explaining the summarization goal and input format, (2) a clear instruction to generate a general claim and distinct reasons with a maximum length of 10 tokens, (3) step-by-step guidelines for how the LLM should transform the input (i.e., infer

³<https://huggingface.co/roberta-large-mnli>

⁴Details in Supplementary Material

⁵<https://platform.openai.com/docs/guides/prompt-engineering>

a concise claim from the claim list, then derive specific, supporting ground key points from each reason cluster).

Experiments & Results

Datasets and Experiment Setting

We evaluate our framework on ConvoSumm (Fabbri et al. 2021), an abstractive conversation summarization corpus with diverse conversation datasets (domains), but specifically focus on three datasets: **NYT** (New York Times news comments), **Reddit** (discussion forums), and **Stack** (Stack-Exchange community question answering), since they are most relevant to the argument-rich discussions targeted by ARQUSUMM. Each instance includes user comments from multi-turn threads, and a crowd-sourced abstractive summary capturing diverse viewpoints. We use only the test split for evaluation, which contains 250 examples per dataset.

We experimented ARQUSUMM with GPT-4.1, and Mistral-7B⁶ as backbone LLMs. For our LLM-based entailment scoring, we sample 5 times to estimate the weighted summation of entailment score of a given claim or reason pair. Note that for runtime and cost feasibility, for all clustering stages, we batched and performed LLM-based Entailment scoring in an one-to-many manner (e.g., scoring 1 claim vs a list of other claims/reasons per prompt) instead of pairwise. We set the clustering threshold $\tau = 3$ based on empirical inspection of the cluster quality.

Baselines

We benchmark ARQUSUMM against a various baselines, covering existing conversation summarization and quantitative summarization (KPA) framework.

ConvoSumm A conversation summarization framework that integrates argument mining with abstractive summarization (Fabbri et al. 2021). Sentences from a comment were first classified as claims or premises, i.e., reasons, before being mapped for relations within and across comments using a RoBERTa (Liu et al. 2019) model finetuned on MNLI entailment⁷ to construct an argument graph. Graph-based information are then linearized as input to a BART-large (Lewis et al. 2020) abstractive summarization model fine-tuned for graph-to-text generation.

GPT-4.1-ICL We few-shot prompt (with two in-context examples) a GPT-4.1 model, as an end-to-end solution, to directly process a list of argument propositions and output a structured summary. The prompt adopts the Chain-of-Thoughts strategy, which guides and elicits the LLM to generate output with four reasoning steps: (1) grouping equivalent claims, (2) aggregating associated reasons, (3) clustering similar justifications, and (4) summarizing each claim with its reason clusters.

QAF An adapted version of the quantitative argumentation framework (QAF) of Gupta, Zuckerman, and O’Connor (2024) for the ARQUSUMM task, which basically clusters argument propositions without distinguishing their roles

(e.g., claim or reason). It follows four stages: (1) extract propositions from comments via LLM prompting, (2) embed and cluster them using Sentence-BERT and DP-means, (3) infer directed edges to recover reason-to-claim links for hypergraph construction, and (4) generate structured summaries from clusters in the first two hypergraph levels.

ConvPAKPA/ConvRKPA Adapted versions of the KPA review summarization systems PAKPA (Tang et al. 2024) and RKPA (Bar-Haim et al. 2021) for conversation summarization, using extracted argument propositions instead of comment sentences. **ConvPAKPA** identifies aspects and associated sentiment, clusters by aspect–sentiment pairs, and prompts LLMs to generate aspect-specific key points. **ConvRKPA** uses a quality ranking model to select KP candidates, then matches propositions using a KP Matching model (Bar-Haim et al. 2020b).

We conducted comprehensive evaluation of our ARQUSUMM framework along the dimensions of argument quality, textual quality of arguments, as well as quantification accuracy. In addition to leveraging LLMs for automatic evaluation, we also employ human evaluation. We organised the results into sections based on research questions.

Summary	Baseline	CV	FF	RD	VL	SN	IN	SA	HF
Structured	ARQUSUMM	26.51	25.34	21.39	31.85	37.41	31.95	23.19	54.46
	GPT-4.1-ICL	22.17	24.54	20.18	30.53	24.43	24.05	21.24	
Textual	QAF	13.60	16.74	18.53	13.23	15.23	20.36	18.94	
	ConvoSumm	13.61	14.49	16.56	12.83	07.73	11.20	13.10	36.75
KP	ConvPAKPA	14.83	11.46	14.36	06.98	10.43	09.02	13.47	08.83
	ConvRKPA	09.28	07.44	08.98	04.58	04.76	03.43	10.06	

Table 1: Human evaluation of summary’s information quality. Reported are the Bradley Terry scores of 8 dimensions, from left to right, COVERAGE, FAITHFULNESS and REDUNDANCY, VALIDITY, SENTIMENT, INFORMATIVENESS, SINGLEASPECT, HELPFULNESS.

RQ1: Is the argument structure in the summaries more helpful to users?

Settings We manually evaluate both *the utility of the argument structure* in the summaries, and also the summary’s *information quality* of different baselines, using a set of 8 evalation dimension. While information quality is adopted from the 7 different dimensions, (e.g., FAITHFULNESS, COVERAGE) defined by previous KPA studies (Kapadnis et al. 2021) (see Appendix), we additionally define HELPFULNESS to evaluate “ how helpful are the viewpoints organized and presented in the summary?”, specifically comparing our new form of claim-reason structured summary over the existing textual and KP summaries. Note that to measure HELPFULNESS, we select summaries generated by the latest work from each form (e.g., ARQUSUMM, ConvoSumm, and ConvPAKPA) to represent the 3 forms.

To perform this evaluation, we hire workers from Amazon Mechanical Turk (MTurk) to conduct pairwise comparisons of KPs from different systems⁸. Each comparison involved choosing the better one from two summaries,

⁶<https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3>

⁷<https://huggingface.co/FacebookAI/roberta-large-mnli>

⁸We ensure inter-annotator consistency by selecting annotators with pairwise Cohen’s Kappa ≥ 0.1

each taken from a different system. Using the Bradley-Terry model (Friedman et al. 2021), we calculated rankings from these comparisons among the models.

Results From Table 1, overall, ARQUSUMM achieves consistent and up to 7.86 times improvements on all 7 dimensions, and are notably higher on COVERAGE, VALIDITY, SENTIMENT and INFORMATIVENESS. In addition, ConvPAKPA achieves a slightly better score in SENTIMENT and SINGLEASPECT than Convosumm thanks to its aspect-sentiment-based clustering approach to capture arguments.

Table 1 further examines the HELPFULNESS of the claim-reason structured summary. Overall, the claim-reason structured summary, produced by ARQUSUMM (ours), GPT-4-ICL and QAF, yields up to 7 times more comprehensive and useful viewpoint presentation than other baselines. In fact, while the traditional textual summary form (of Convosumm) already grounded generation with argument information for diversity, it failed to attach convincing evidence along viewpoints, nor presenting them in a logical manner. Notably, the KP summary form, by producing a long and flat list of KPs, makes least sense of comprehension and usefulness as overlapping KPs (at different granularity) being scattered across the list inattentively.

RQ2: How well the generated summaries represent arguments from the corpus?

Settings This evaluation ignores argument structure and instead assesses the textual quality of arguments in generated summaries against the gold summaries. We aggregate all claims and reasons from each structured summary and reference summary as argument sets. Lexical similarity is computed using maximum ROUGE scores between generated and reference arguments. Then, following Li et al. (2023), we calculate soft-Precision/Recall/F1 (denoted as sP , sR and $sF1$) to evaluate the *semantic* similarity between individual generated argument and reference argument. While sP finds the reference argument with the highest similarity score for each generated argument, sR is vice-versa, and ($sF1$) is the harmonic mean between sP and sR .

$$sP = \frac{1}{n} \times \sum_{\alpha_i \in \mathcal{A}} \max_{\beta_j \in \mathcal{B}} f(\alpha_i, \beta_j) \quad (2)$$

$$sR = \frac{1}{m} \times \sum_{\beta_j \in \mathcal{B}} \max_{\alpha_i \in \mathcal{A}} f(\alpha_i, \beta_j) \quad (3)$$

where f computes similarities between two individual key points, \mathcal{A} , \mathcal{B} is the set of generated and reference KPs and $n = |\mathcal{A}|$ and $m = |\mathcal{B}|$, respectively. We use state-of-the-art semantic similarity metrics BLEURT (Sellam, Das, and Parikh 2020) and BERTScore (Zhang et al. 2020b). Additionally, we further measure the utility of argumentation theory in conversation summarization, by setting up a regular PAKPA that directly accepts user comments as input.

Results Table 2 reports the lexical and semantic quality of generated argument in the summaries, with Reddit being the most challenging domain due to its informal language. Overall, ARQUSUMM consistently outperforms all baselines, achieving up to 3.71 times higher lexical similarity

and a 0.22-point gain in semantic quality, thanks to disentangling the claim-reason clustering process in our argument-aware clustering. This ultimately prevents reasons from being mixed within claim clusters, ensuring more specific and diverse justifications. In contrast, QAF, which clusters propositions without role distinction, ranks second-lowest with weaker coherence (e.g., $sF1 = 0.18$ on Reddit). Nevertheless, our use of entailment scoring in ARQUSUMM, rather than simple semantic similarity (e.g., QAF, also helps obtain a clearer distinction of subtle attitudes and stances, contributing to its superior sR (0.36 vs. 0.27 BLUERT $sF1$ on NYT). Importantly, utilizing lightweight LLMs (e.g., Mistral) as backbone for ARQUSUMM does not substantially degrades the textual quality due to our rigorous claim-reason clustering process.

Among the weaker baselines, KPA-based models (ConvPAKPA and ConvRKPA) perform poorly as they were not designed to process and summarize argument structure between comments’ claims and reasons. Nevertheless, ConvPAKPA outperforms QAF in Reddit and NYT, as its separate generation of claim and reason KPs, which help capture viewpoints more effectively, reflected in higher sR scores (0.34 vs. 0.27 BLUERT $sF1$ on NYT)

Finally, although GPT-4.1-ICL is a strong generative baseline, it underperforms ARQUSUMM in both lexical quality (up to 19.3%) and semantic quality (up to 22.9%). This is primarily due to hallucinations and the challenges LLMs face with long-context reasoning, where multi-step reasoning over complex propositions can make its performance less robust compared to ARQUSUMM.

RQ3: How well does the model match between the comments and the generated claims or reasons?

Settings We evaluate how accurately each framework clusters original claims and reasons from input comments and matches them with those in the generated summary. We extend Bar-Haim et al. (2021) to measure both *precision* (correctness of predicted matches) and *recall* (coverage of ground-truth matches), by prompting gpt-4-o-mini to annotate pairwise *match/non-match* between generated and original claims or reasons. Results are reported at three levels: claim, reason, and claim-reason. The claim-reason level offers a stricter evaluation of the argument structure, requiring both a reason and its parent claim to be correctly matched to their respective summarized counterparts, thus evaluating structural consistency. Empirical validation shows gpt-4-o-mini annotations highly correlated with MTurk workers’ judgement (Pearson’s $r = 0.647$).

Results Table 3 reports how well different systems cluster claims and reasons in the final summary, evaluated at the claim, reason, and claim-reason levels. ARQUSUMM mostly achieves the highest precision, recall, and F1 across domains (e.g., 0.563 vs 0.142 Reason-Level F1 for Reddit)

Across all three domains, we also observe that ARQUSUMM and GPT-4.1-ICL achieve highly stable, and sometimes higher, performance at the reason level compared to the claim level, largely due to their argument-aware clustering strategy. By first clustering claims and then aggre-

gating their child reasons for reason-level clustering, these frameworks create more coherent reason groups. As a result, they achieve stronger structural consistency, which in turn boost the performance at the claim-reason level. (e.g., 0.728 F1 for ARQUSUMM on Stack) We also notice that limited computation of lightweight LLMs (e.g., Mistral) makes them score and cluster claims/reasons not as accurate as GPT-4.1 in ARQUSUMM, and even GPT-4.1-ICL.

In contrast, QAF, without disentangling the role of claims and reasons for clustering nor adopting hierarchical clustering, shows much lower reason-level performance, especially in the NYT domain (e.g., 0.168 F1 vs. 0.512 F1 of ARQUSUMM). In fact, with higher volume and complexity of arguments across NYT comments, reasons were more likely to be mixed in the claim cluster and therefore cannot contribute meaningfully to their truly expected reason clusters.

Similarly, KPA-based approaches also underperform because they are not originally designed to model argument structure. However, ConvPAKPA still achieves better clustering performance than QAF thanks to separately processing claim and reason during clustering to avoid mixing.

RQ4: How well and reasonable does the reason support its parent claim in an argument?

Settings We further evaluate the convincingness of reasons supporting its parent claim along the structure in our generated summary leveraging claim verification task in fact checking. Given a claim and every of its associated reason in the summary, we prompt gpt-4.1-mini to annotate whether the reason *supports* or *refutes* the claim. We then compute the precision of the *support* label, which is the proportion of reasons judged as valid evidence for their associated claims.

Results Table 4 shows that ARQUSUMM consistently achieves the highest reason-claim support across all domains. (e.g., 0.828 for ARQUSUMM vs 0.738 for GPT-4.1-ICL on Stack). This highlight the effectiveness of ARQUSUMM’s entailment-based clustering and structured summary generation in maintaining logical coherence between claims and reasons. GPT-4.1-ICL, while performing better than QAF, surpassed by ARQUSUMM due to its one-step generation strategy, which can produce plausible but less rigorously justified reasons. However, using lightweight LLMs for ARQUSUMM cannot surpass GPT-4.1-ICL as these models cannot perform entailment-based scoring and clustering of claims/reasons as accurate as GPT-4.1. QAF performs the weakest (0.568 on NYT), as its role-agnostic approach to clustering often leads to mismatched or loosely related reasons being associated with claims.

Conclusion

In this paper, we introduced ARQUSUMM, a novel task and framework of argument-aware quantitative summarization for online discussions which generate structured summaries composed of claims and their supporting reasons. Different from previous works, our approach leverages entailment-based clustering to disentangle claims and reasons, followed by hierarchical aggregation to preserve both specificity and diversity of viewpoints. We further designed a structured

	ROUGE			BERTScore			BLEURT		
	R-1	R-2	R-L	sP	sR	sF1	sP	sR	sF1
Reddit									
ARQUSUMM (GPT-4.1)	0.368	0.167	0.342	0.19	0.31	0.24	0.28	0.34	0.31
ARQUSUMM (Mistral)	0.357	0.151	0.335	0.19	0.28	0.23	0.32	0.33	0.33
GPT-4.1-ICL	0.324	0.138	0.306	0.17	0.25	0.20	0.28	0.33	0.30
ConvоТumm	0.337	0.128	0.312	0.26	0.22	0.24	0.36	0.31	0.33
QAF	0.239	0.068	0.219	0.19	0.18	0.18	0.28	0.26	0.27
ConvPAKPA	0.282	0.070	0.257	0.23	0.27	0.24	0.30	0.26	0.28
ConvRKPA	0.177	0.045	0.169	0.18	0.13	0.15	0.26	0.22	0.24
Stack									
ARQUSUMM (GPT-4.1)	0.480	0.275	0.447	0.24	0.35	0.28	0.38	0.42	0.40
ARQUSUMM (Mistral)	0.466	0.250	0.412	0.25	0.28	0.26	0.41	0.39	0.40
GPT-4.1-ICL	0.439	0.222	0.404	0.21	0.27	0.24	0.35	0.40	0.38
ConvоТumm	0.412	0.191	0.364	0.31	0.23	0.27	0.43	0.34	0.38
QAF	0.402	0.173	0.366	0.26	0.22	0.24	0.41	0.36	0.38
ConvPAKPA	0.321	0.104	0.280	0.24	0.25	0.24	0.38	0.30	0.34
ConvRKPA	0.256	0.084	0.242	0.23	0.13	0.17	0.34	0.26	0.29
NYT									
ARQUSUMM (GPT-4.1)	0.444	0.235	0.410	0.20	0.36	0.26	0.32	0.31	0.36
ARQUSUMM (Mistral)	0.428	0.213	0.391	0.21	0.32	0.25	0.33	0.38	0.35
GPT-4.1-ICL	0.408	0.196	0.378	0.22	0.31	0.25	0.32	0.40	0.35
ConvоТumm	0.371	0.165	0.342	0.30	0.23	0.26	0.39	0.30	0.34
QAF	0.356	0.144	0.331	0.24	0.28	0.26	0.33	0.27	0.27
ConvPAKPA	0.365	0.137	0.334	0.25	0.30	0.27	0.36	0.32	0.34
ConvRKPA	0.270	0.083	0.250	0.21	0.20	0.21	0.28	0.25	0.27

Table 2: General textual quality of generated claim and reason KPs from the summary.

	Claim Level			Reason Level			Claim-Reason Level		
	P	R	F1	P	R	F1	P	R	F1
Reddit									
ARQUSUMM (GPT-4.1)	0.834	0.376	0.518	0.813	0.430	0.563	0.647	0.788	0.711
GPT-4.1-ICL	0.771	0.251	0.379	0.626	0.375	0.469	0.513	0.701	0.592
ARQUSUMM (Mistral)	0.747	0.169	0.276	0.590	0.387	0.468	0.516	0.604	0.557
QAF	0.683	0.157	0.255	0.581	0.265	0.364	0.470	0.581	0.519
ConvPAKPA	0.405	0.375	0.389	0.316	0.603	0.415	—	—	—
ConvRKPA	0.265	0.150	0.191	0.176	0.119	0.142	—	—	—
Stack									
ARQUSUMM (GPT-4.1)	0.811	0.403	0.538	0.804	0.369	0.506	0.627	0.867	0.728
GPT-4.1-ICL	0.730	0.147	0.245	0.610	0.319	0.419	0.572	0.626	0.598
ARQUSUMM (Mistral)	0.641	0.141	0.231	0.660	0.301	0.413	0.618	0.528	0.569
QAF	0.618	0.130	0.214	0.750	0.241	0.364	0.417	0.562	0.479
ConvPAKPA	0.680	0.233	0.347	0.357	0.565	0.438	—	—	—
ConvRKPA	0.500	0.125	0.200	0.533	0.072	0.127	—	—	—
NYT									
ARQUSUMM (GPT-4.1)	0.573	0.558	0.566	0.756	0.387	0.512	0.485	0.837	0.614
GPT-4.1-ICL	0.827	0.188	0.307	0.691	0.298	0.416	0.592	0.604	0.598
ARQUSUMM (Mistral)	0.523	0.212	0.302	0.656	0.282	0.394	0.365	0.627	0.461
QAF	0.517	0.144	0.225	0.367	0.109	0.168	0.250	0.050	0.080
ConvPAKPA	0.568	0.326	0.414	0.427	0.585	0.494	—	—	—
ConvRKPA	0.431	0.264	0.327	0.257	0.111	0.155	—	—	—

Table 3: Comment matching correctness of claim and reason KPs in the generated summary, measured at different level. Claim-Reason Level not applicable for ConvPAKPA and ConvRKPA as they were not designed to process and summarize argument structure between claims and reasons.

	Reddit	Stack	NYT
ARQUSUMM (GPT-4.1)	0.804	0.828	0.764
GPT-4.1-ICL	0.687	0.738	0.612
ARQUSUMM (Mistral)	0.669	0.710	0.594
QAF	0.652	0.675	0.568

Table 4: *Precision* of reason-claim support (convincingness) from the generated summary. Applicable only to baselines outputting claim-reason structured summary.

generation strategy to ensure logical alignment between claims and reasons, addressing the limitations of existing textual summarization methods. Experiments on multiple forms of conversation social comments demonstrate that ARQUSUMM consistently delivers higher lexical and semantic quality, better clustering performance, and more convincing reasoning compared to baselines.

Acknowledgement

This research is supported in part by the Australian Research Council Discovery Project **DP200101441**.

References

- Bar-Haim, R.; Eden, L.; Friedman, R.; Kantor, Y.; Lahav, D.; and Slonim, N. 2020a. From Arguments to Key Points: Towards Automatic Argument Summarization. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4029–4039. Online: Association for Computational Linguistics.
- Bar-Haim, R.; Eden, L.; Kantor, Y.; Friedman, R.; and Slonim, N. 2021. Every Bite Is an Experience: Key Point Analysis of Business Reviews. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *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)*, 3376–3386. Online: Association for Computational Linguistics.
- Bar-Haim, R.; Kantor, Y.; Eden, L.; Friedman, R.; Lahav, D.; and Slonim, N. 2020b. Quantitative argument summarization and beyond: Cross-domain key point analysis. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 39–49. Online: Association for Computational Linguistics.
- Barker, E.; and Gaizauskas, R. 2016. Summarizing Multi-Party Argumentative Conversations in Reader Comment on News. In Reed, C., ed., *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, 12–20. Berlin, Germany: Association for Computational Linguistics.
- Barker, E.; Paramita, M. L.; Aker, A.; Kurtic, E.; Hepple, M.; and Gaizauskas, R. 2016. The SENSEI Annotated Corpus: Human Summaries of Reader Comment Conversations in On-line News. In Fernandez, R.; Minker, W.; Carenini, G.; Higashinaka, R.; Artstein, R.; and Gainer, A., eds., *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 42–52. Los Angeles: Association for Computational Linguistics.
- Cattan, A.; Eden, L.; Kantor, Y.; and Bar-Haim, R. 2023. From Key Points to Key Point Hierarchy: Structured and Expressive Opinion Summarization. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 912–928. Toronto, Canada: Association for Computational Linguistics.
- Chen, J.; and Yang, D. 2021. Structure-Aware Abstractive Conversation Summarization via Discourse and Action Graphs. In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tur, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1380–1391. Online: Association for Computational Linguistics.
- Fabbri, A.; Rahman, F.; Rizvi, I.; Wang, B.; Li, H.; Mehdad, Y.; and Radev, D. 2021. ConvoSumm: Conversation Summarization Benchmark and Improved Abstractive Summarization with Argument Mining. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *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)*, 6866–6880. Online: Association for Computational Linguistics.
- Freeman, J. B. 1991. *Dialectics and the Macrostructure of Arguments: A Theory of Argument Structure*. 10. Walter de Gruyter.
- Friedman, R.; Dankin, L.; Hou, Y.; Aharonov, R.; Katz, Y.; and Slonim, N. 2021. Overview of the 2021 Key Point Analysis Shared Task. In Al-Khatib, K.; Hou, Y.; and Stede, M., eds., *Proceedings of the 8th Workshop on Argument Mining*, 154–164. Punta Cana, Dominican Republic: Association for Computational Linguistics.
- Geffet, M.; and Dagan, I. 2004. Feature Vector Quality and Distributional Similarity. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, 247–253. Geneva, Switzerland: COLING.
- Geffet, M.; and Dagan, I. 2005. The Distributional Inclusion Hypotheses and Lexical Entailment. In Knight, K.; Ng, H. T.; and Oflazer, K., eds., *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL '05)*, 107–114. Ann Arbor, Michigan: Association for Computational Linguistics.
- Gliwa, B.; Mochol, I.; Bieseck, M.; and Wawer, A. 2019. SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization. In Wang, L.; Cheung, J. C. K.; Carenini, G.; and Liu, F., eds., *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, 70–79. Hong Kong, China: Association for Computational Linguistics.
- Gupta, A.; Zuckerman, E.; and O'Connor, B. 2024. Harnessing Toulmin's theory for zero-shot argument explication. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10259–10276. Bangkok, Thailand: Association for Computational Linguistics.
- Kapadnis, M.; Patnaik, S.; Panigrahi, S.; Madhavan, V.; and Nandy, A. 2021. Team Enigma at ArgMining-EMNLP 2021: Leveraging Pre-trained Language Models for Key Point Matching. In Al-Khatib, K.; Hou, Y.; and Stede, M., eds., *Proceedings of the 8th Workshop on Argument Mining*, 200–205. Punta Cana, Dominican Republic: Association for Computational Linguistics.
- Lenz, M.; Sahitaj, P.; Kallenberg, S.; Coors, C.; Dumani, L.; Schenkel, R.; and Bergmann, R. 2020. Towards an argument mining pipeline transforming texts to argument graphs. In *Computational Models of Argument*, 263–270. IOS Press.
- Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault,

- J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880. Online: Association for Computational Linguistics.
- Li, H.; Schlegel, V.; Batista-Navarro, R.; and Nenadic, G. 2023. Do You Hear The People Sing? Key Point Analysis via Iterative Clustering and Abstractive Summarisation. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14064–14080. Toronto, Canada: Association for Computational Linguistics.
- Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2511–2522. Singapore: Association for Computational Linguistics.
- Liu, Y.; and Lapata, M. 2019. Text Summarization with Pre-trained Encoders. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3730–3740. Hong Kong, China: Association for Computational Linguistics.
- Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Misra, A.; Anand, P.; Tree, J. E. F.; and Walker, M. 2017. Using summarization to discover argument facets in online ideological dialog. *arXiv preprint arXiv:1709.00662*.
- Sellam, T.; Das, D.; and Parikh, A. 2020. BLEURT: Learning Robust Metrics for Text Generation. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7881–7892. Online: Association for Computational Linguistics.
- Stab, C.; and Gurevych, I. 2014. Identifying Argumentative Discourse Structures in Persuasive Essays. In Moschitti, A.; Pang, B.; and Daelemans, W., eds., *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 46–56. Doha, Qatar: Association for Computational Linguistics.
- Syed, S.; Ziegenbein, T.; Heinisch, P.; Wachsmuth, H.; and Potthast, M. 2023. Frame-oriented Summarization of Argumentative Discussions. In Stoyanchev, S.; Joty, S.; Schlangen, D.; Dusek, O.; Kennington, C.; and Alikhani, M., eds., *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 114–129. Prague, Czechia: Association for Computational Linguistics.
- Tang, A.; Zhang, X.; and Dinh, M. 2024. Aspect-based Key Point Analysis for Quantitative Summarization of Reviews. In *18th Conference of the European Chapter of the Association for Computational Linguistics*.
- Tang, A.; Zhang, X.; Dinh, M.; and Cambria, E. 2024. Prompted Aspect Key Point Analysis for Quantitative Re-view Summarization. In Ku, L.-W.; Martins, A.; and Sriku-mar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10691–10708. Bangkok, Thailand: Association for Computational Linguistics.
- Toulmin, S. 1958. The uses of argument cambridge university press. Cambridge, UK.
- Völske, M.; Potthast, M.; Syed, S.; and Stein, B. 2017. TL;DR: Mining Reddit to Learn Automatic Summarization. In Wang, L.; Cheung, J. C. K.; Carenini, G.; and Liu, F., eds., *Proceedings of the Workshop on New Frontiers in Summarization*, 59–63. Copenhagen, Denmark: Association for Computational Linguistics.
- Walton, D. N. 1996. *Argumentation Schemes for Presumptive Reasoning*. Psychology Press.
- Zhang, H.; Wang, S.; Chen, T.-H.; and Hassan, A. E. 2019. Reading answers on stack overflow: Not enough! *IEEE Transactions on Software Engineering*, 47(11): 2520–2533.
- Zhang, J.; Zhao, Y.; Saleh, M.; and Liu, P. 2020a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, 11328–11339. PMLR.
- Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020b. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.