Argument Summarization: A Unified Task Definition, Taxonomy, and Survey

Zhiqiang Shi
Department of Informatics
King's College London
zhiqiang.shi@kcl.ac.uk

Oana Cocarascu
Department of Informatics
King's College London
oana.cocarascu@kcl.ac.uk

Abstract

Argument summarization seeks to generate concise representations of argumentative content while preserving argumentative meaning, yet the field remains conceptually fragmented. In this survey, we provide a unified task definition of argument summarization that encompasses both argumentative text summarization and key point analysis, and introduce a taxonomy along input structure, output representation, and the role of argumentative structure. Using this framework, we analyze the evolution of the field and systematically review existing datasets, methods, and evaluation practices, revealing misalignment that hinders comparability and progress. We further situate argument summarization within the broader context of generic text summarization, identifying practical limitations and intrinsic challenges. Finally, we outline promising future research directions.

1 Introduction

Argument summarization aims to generate concise representations of argumentative content while preserving its underlying argumentative meaning. Humans routinely summarize debates, legal opinions, scientific controversies, and online discussions by abstracting central claims, grouping supporting and opposing reasons, and making relations between viewpoints explicit. Replicating this ability computationally has therefore attracted interest at the intersection of text summarization, computational argumentation, and natural language generation.

However, argument summarization has emerged as a fragmented research area. The same term has been used to describe different problem formulations, input assumptions, output representations, evaluation practices (Bar-Haim et al., 2020a; Li et al., 2025; Tang et al., 2026). Thus, findings are difficult to compare across studies, modeling decisions are often insufficiently justified, and reported improvements may reflect properties of particular tasks or evaluation settings rather than genuine advances in summarizing argumentative structure.

A key source of fragmentation is that argument summarization encompasses at least two tasks. The first, which we refer to as Argumentative Text Summarization (ATS) (Li et al., 2025), summarizes a single argumentative or semi-argumentative document (e.g. a legal decision, editorial, discussion thread) into a coherent summary that captures both content and argumentative flow. Here, the input is an internally coherent discourse with an implicit or explicit argument structure, and the output is typically a free-form abstractive summary that reflects this structure. The second originates from Key Point Analysis (KPA) (Bar-Haim et al., 2020a) and takes as input multiple independent arguments, often short user-generated statements, with the goal to produce a set of key points representing the main perspectives in the argument set, along with links mapping arguments to the key points they support. This setup is more closely aligned with computational argumentation, clustering, and semantic matching than to traditional summarization.

We argue that many of the challenges in this area stem from treating these different tasks as interchangeable. Tasks with different assumptions regarding inputs, outputs, and objectives are often labeled as argument summarization (Li et al., 2025; Altemeyer et al., 2025). Moreover, models are frequently evaluated with metrics that are misaligned with the task objectives, making results difficult to interpret and giving a misleading view of progress.

However, at a broader level, these tasks share the same goal: generating concise representations of argumentative content while preserving argumentative meaning. Building on this shared objective while maintaining clear distinctions between task formulations, we propose a unified definition of argument summarization and introduce a taxonomy along three dimensions: input structure, output representation, and the role of argumentative structure.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 307–319
November 6-10, 2026 ©2026 Association for Computational Linguistics

Under our unified definition and taxonomy, we first examine the temporal evolution of the field highlighting how task formulations, modeling paradigms, and research directions have developed over time. We then systematically review methods, datasets, and evaluation metrics, showing how their misalignments have led to fragmentation, while also situating them within a unified conceptual space. Finally, we position argument summarization within the context of broader NLP tasks, emphasizing its differences from generic text summarization, the unique challenges posed by argumentative structure, and future research opportunities.

Our contributions are: 1) We propose a **unified definition of argument summarization** grounded in the shared goal of producing concise representations of argumentative content, encompassing ATS and KPA; 2) We introduce a three-dimensional **taxonomy of task formulations**, covering input structure, output representation, and the role of argumentative structure, that provides a principled framework for analyzing and comparing existing approaches for argument summarization; 3) We trace the field's **temporal evolution** across task formulations and modeling paradigms; 4) We **systematically analyze methods, data, and evaluation metrics** using the proposed taxonomy to explain fragmentation and unify existing work; 5) We situate argument summarization within the **broader NLP landscape**, highlighting key differences from generic text summarization, identifying unique challenges, and outlining future directions.

2 Task Definition

Argumentative text summarization (ATS): ATS aims to generate a concise and coherent summary from a single argumentative or semi-argumentative text, while preserving the argumentative structure. The input is a continuous discourse with an internal argument flow, and the output is a natural language summary that preserves both the semantic content and the argumentative structure of the input text.

Key point analysis (KPA): KPA frames argument summarization as abstracting a set of independent arguments into a small set of concise, representative key points, along with mappings from each argument to the supported key point(s). Unlike ATS, which summarizes a single continuous argumentative discourse, KPA operates on collections of short, stand-alone arguments (e.g. user-generated comments, debate posts, crowd-sourced opinions).

A unified definition: While the two definitions differ, both aim to compress argumentative content while preserving its argumentative meaning. Therefore, in this paper, we provide a unified definition for argument summarization: *Argument summarization refers to the task of generating concise representations of argumentative content that preserve its underlying argumentative meaning, where the input may be a single argumentative discourse, a set of independent arguments or other forms of argumentative content, and the output may take the form of a coherent textual summary, a set of key points, or other structured representations.*

Next, we introduce a taxonomy that highlights the key dimensions along which existing argument summarization approaches differ, providing a common framework for analyzing methods, datasets, and evaluation practices.

3 A Unified Taxonomy for Argument Summarization

Argument summarization research spans diverse task formulations, modeling choices, output formats, and evaluation practices, reflecting the variety of argumentative settings and applications, but resulting in a fragmented research landscape. To better understand and compare underlying assumptions about inputs, outputs, and evaluation, we propose a taxonomy that organizes argument summarization along three orthogonal dimensions: input structure, output representation, and the role of argumentative structure. These dimensions capture the main design choices that distinguish task formulations and help explain why methods, datasets, and evaluation metrics are often misaligned.

Input Structure. Argument summarization systems vary in input type and coherence: *Single argumentative documents* (long texts, e.g., legal opinions or editorials with internal argumentative flow), *Multi-turn discussions* (online conversations, debates where argumentation is distributed across speakers); *Independent short arguments* (sets of stand-alone comments or statements, typically in KPA); *Partially structured evidence sets* (claim-evidence pairs, annotated argument units). This dimension captures the coherence, granularity, and argument density of the input.

Output Representation. Systems vary in how they represent summarized argumentative content: *Abstractive summaries* (continuous natural-language

summaries, common in ATS); *Key points* (short propositions representing central perspectives, typically in KPA); *Structured outputs* (claim-reason chains, issue-reason-conclusion frames, argument graphs); *Hybrid outputs* (textual summaries with argument roles or viewpoint structure). This dimension reflects how much argumentative information is made explicit in the output, with direct implications for what aspects of argumentative meaning are preserved and how summaries are evaluated.

Structure Role. Argumentative structure may be incorporated differently during modeling: *Structure-agnostic* (no argument information is used, e.g., generic summarization); *Structure-aware* (argument roles, stance, similarity, or argument graphs are used as features or intermediate signals, but the final output does not encode argument structure); *Structure-preserving* (systems generate summaries that maintain the underlying argumentative relations such as claim-reason links and issue-reason-conclusion flow, and can be expressed in explicit structured form or as coherent abstractive text). This dimension explains differences in methodology and the substantial variation in evaluation across studies.

We use this taxonomy throughout the remainder of the survey to organize methods, analyze datasets, and examine evaluation practices, and to explain how mismatches across these dimensions have contributed to the perceived fragmentation of the field.

4 Survey Methodology

Literature was collected through searches in ACL Anthology, Google Scholar, and Semantic Scholar, supplemented by backward and forward citation tracing from influential works in argument summarization, key point analysis, argument mining, and text summarization. We include studies whose primary objective is the summarization of argumentative content, including Argumentative Text Summarization (ATS), Key Point Analysis (KPA), and related datasets, methods, and evaluation approaches. We additionally include a limited number of representative text summarization studies where necessary to contextualize argument summarization within the broader NLP landscape and to facilitate comparisons of datasets, methods, and evaluation practices.

5 Temporal Evolution of Argument Summarization

We trace the temporal evolution of argument summarization research, highlighting shifts in task formulations, modeling approaches, and evaluation practices. Table 1 summarizes selected works, organized using our proposed taxonomy.

5.1 Key Point Analysis

KPA aims to condense a set of arguments into a small set of concise summaries (i.e. key points). Unlike document-centric summarization, KPA inputs are sets of independent arguments, an assumption that has remained consistent across prior work. Early work that later influenced KPA emerged from research on argument similarity, clustering, and abstraction. These studies addressed related tasks such as extracting salient elements from argumentative discussions (Egan et al., 2016) and identifying arguments expressing the same underlying facet or perspective (Misra et al., 2016; Reimers et al., 2019). Framing-based approaches further organized arguments into non-overlapping frames that capture shared viewpoints (Ajjour et al., 2019). While these works did not define argument summarization as a standalone task, they introduced set-based representations and abstraction mechanisms that later became central to KPA.

KPA was formally defined by Bar-Haim et al. (2020a,b) and marked a conceptual shift by fixing both the input structure (independent arguments) and the output representation (a set of key points). The KPA shared task at ArgMining 2021 (Friedman et al., 2021) further standardized this formulation by dividing it into two sub-tasks: key point generation (KPG) and key point matching (KPM). In KPG, systems generate pro and con key points for a given topic; in KPM, individual arguments are matched to the key points they support. Building on this formalization, a substantial body of work emerged around the shared task. For KPG, both extractive and abstractive methods were proposed: Alshomary et al. (2021) developed an extractive method to select the highest quality arguments as key points candidates, followed by the computation of an importance score to select the final key points, while Kapadnis et al. (2021) proposed an abstractive method where key points candidates are generated by a pre-trained language model (PLM) from the concatenated arguments, and the final key points are selected based on ROUGE scores. For

Task	Works	Input	Output	Structure Role
multi-document abstractive summarization	Alshaina et al. (2017)	multiple documents	abstractive summary based on predicate argument structure	structure-aware
news editorials summarization	Syed et al. (2020)	news editorials	extractive summary with sentences from different evidence types	structure-aware
text summarization using formal argumentation	Sukprapa et al. (2021)	argumentative text	text summary based on argument structure	structure-aware
abstractive conversation summarization with argument mining	Fabbri et al. (2021a)	online conversations	summary incorporating argument graph	structure-aware
legal case summarization	Xu et al. (2021)	legal case decisions	issue-reason-conclusion extractive summary	structure-preserving
legal document summarization	Elaraby and Litman (2022)	legal documents	summary that utilises argument roles	structure-aware
extractive summary of online discussions	Syed et al. (2023)	online discussions	extractive summaries grouped into argumentative frames	structure-aware
abstractive summarization of legal opinions	Elaraby et al. (2023)	legal opinions	argument-aware (issue, reason, conclusion) abstractive summary	structure-preserving
objective argument summarization in search	Ziegenbein et al. (2024)	website pages	objective argument snippets	structure-aware
argument summarization using claim-evidence pairs	Li et al. (2025)	claim-evidence pairs	abstractive summary generated from claim-evidence pairs	structure-preserving
product review summarization	Zhou et al. (2025)	product reviews, aspects, argumentation schemes	aspect-based summary	structure-aware
topic-level contrastive opinion summarization	Shan and Lu (2025)	online discussions	contrastive opinion summary based on argument relations	structure-aware
argument-aware conversation summarization	Tang et al. (2026)	online conversations	claim-reason structured summary	structure-preserving
key point analysis (KPA)	Bar-Haim et al. (2020a); Gurjar et al. (2025)	a set of arguments	a set of key points	structure-aware

Table 1: Representative works in argument summarization organized using the proposed taxonomy.

KPM, the task was formulated as a sentence-pair classification problem (Reimer et al., 2021), a contrastive learning problem (Phan et al., 2021), and an embedding space learning problem where the matching argument-key points pairs are closer than non-matching pairs (Alshomary et al., 2021). As KPA was formalized in the pre-trained LM era, most studies in this period used PLMs such as RoBERTa (Liu et al., 2019).

With the success of large language models (LLMs), KPA research has increasingly used them. Tang et al. (2024a) used LLMs to review KPA to extract aspects and sentiments from comments, cluster those with shared aspects and sentiments, and generate key points for each cluster. Altemeyer et al. (2025) provided a systematic analysis of strategies for integrating LLMs into existing KPA pipelines and examined their impact on evaluation. While LLMs enable more flexible and abstractive KPG, they do not fundamentally alter KPA's core assumptions: inputs remain independent arguments and outputs remain sets of key points. Table 2 summarizes representative KPA works.

5.2 Argumentative Text Summarization

ATS emerged within general text summarization. Early works did not explicitly frame it as argument summarization, but explored how argumentative or discourse structure can improve summaries of argumentative texts. Alshaina et al. (2017) used predicate-argument structures for abstractive summarization across multiple documents. Syed et al. (2020) generated extractive summaries of news editorials using different evidential roles. In these works, argumentative information acted as an auxiliary signal for general summarization.

Subsequent work shifted toward generating summaries that more explicitly reflect the input's argumentative content, though the term argument summarization was not used. Fabbri et al. (2021a) built graphs of issues, viewpoints, and assertions to support abstractive conversation summarization. Elaraby et al. (2023) proposed an argument-aware abstractive summarization framework for legal opinions structured around issues, reasons, and conclusions. These works move from structure-aware modeling toward structure-preserving outputs, particularly in domain-specific settings.

Work	Category	Method
Bar-Haim et al. (2020b)	extractive KPG	KP candidates are selected using heristics and a Quality Scorer, then a Match Scorer and ranking algorithm are used to select the final KPs.
Alshomary et al. (2021)	extractive KPG	Arguments with high quality score are selected as KP candidate, then an important score is computed to select the final KPs.
Kapadnis et al. (2021)	abstractive KPG	A pre-trained LM is used to generate KP candidates given the concatenated arguments, ROUGE score is used to select the final KPs.
Li et al. (2023)	abstractive KPG	Arguments are clustered into different groups, and FLAN-T5 is then used to generate KPs for each clusters.
Altemeyer et al. (2025)	abstractive KPG	LLM end-to-end summarization, an LLM is asked to generate KPs given the set of arguments and topic.
Alshomary et al. (2021)	KPM	An embedding space is learnt where the matching argument-KP pairs are closer than non-matching pairs.
Reimer et al. (2021)	KPM	Fine-tuned RoBERTa for sentence-pair classification for argument-KP pairs.
Phan et al. (2021)	KPM	Embeddings for argument and KP pairs are obtained using RoBERTa, then contrastive learning is applied by grouping matching and non-matching arguments for the KP.
Tang et al. (2024a)	KPA	LLMs extract aspects and sentiment from comments, comments sharing same aspects and sentiments are clustered, and KPs are generated for these clusters.

Table 2: Representative works for key point analysis.

More recently, ATS has been expanded through dataset-driven task definitions. Li et al. (2024a) proposed a dataset for end-to-end argument summarization and evaluation, covering evidence detection and ranking, argument summarization, and summary quality evaluation, where summaries are formed by selecting and organizing supporting evidence into a narrative. Later, Li et al. (2025) framed argument summarization as generating abstractive summaries from claim-evidence pairs.

Unlike KPA, ATS lacks a single, stable task definition, with studies varying in input formats, output representations, and interpretations of how argumentative content should be preserved (see Table 1). For instance, Li et al. (2025) framed argument summarization as generating an abstractive summary from claim-evidence pairs, Ziegenbein et al. (2024) produced objective argumentative snippets for website pages, and Tang et al. (2026) generated summaries that encode claim-reason structures in online discussions. This heterogeneity has led to diverse ATS formulations that are hard to compare directly, contributing to the conceptual fragmentation of argument summarization research.

6 Datasets

Dataset development has been central in defining and evaluating argument summarization. Without a unified task formulation, existing datasets implicitly encode different assumptions about inputs, outputs, and argumentative content preservation. Thus, examining these datasets helps explain the divergence between ATS and KPA, and the sources of fragmentation in the field. Table 3 shows representative datasets for argument summarization.

Bar-Haim et al. (2020a) introduced ArgKP, the first dataset for KPA, with arguments written by crowdsource workers and key points by experts given only the topic. Friedman et al. (2021) extended ArgKP and released ArgKP21 for the shared task. Nearly all subsequent KPA works evaluate their methods on ArgKP21. However, as noted in Gurjar et al. (2025), ArgKP21 has several limitations: the input arguments are short sentences, independent from each other. To overcome these, Gurjar et al. (2025) proposed ArgCMV where arguments are multi-turn, long-context online debates from Reddit ChangeMyView. Different from ArgKP, ArgCMV used LLMs for annotation, followed by human validation on a data subset. Other datasets (e.g. Perspectrum (Chen et al., 2019)) can be adapted for KPA, although not designed for it. These datasets reflect the core KPA assumptions of independent inputs and set-based abstraction, corresponding to the input structure and output representation dimensions of our taxonomy.

For ATS, the lack of a unified task definition led to different input and output formats, making most datasets domain- and format-specific. ConvoSumm (Fabbri et al., 2021a) has four datasets of online discussions across domains, with discussions as inputs and abstractive summaries as outputs created using an issues-viewpoints-assertions framework. WebisEditorialSum-2020 (Syed et al., 2020) has extractive summaries with sentences from different evidence types for news editorials. Li et al. (2024a) proposed a multitask dataset for end-to-end argument summarization, with claim-evidence pairs as inputs and abstractive summaries as outputs.

Many works rely on existing text summarization

Dataset	Domain	Size	Notes
DebateSum (Roush and Balaji, 2020)	debate evidence	187,386	Each row contains an evidence, an extractive summary of it, and an argument made with it.
OpenDebateEvidence (Roush et al., 2024)	debate evidence	3.5M	Contains evidence, extractive and abstractive summary for them and metadata.
ConvoSumm (Fabbri et al., 2021a)	online discussions	contains 4 datasets	abstractive summary created using issues-viewpoints-assertions framework.
WebisEditorialSum-2020 (Syed et al., 2020)	news editorials	266 editorials (each has 5 summaries)	extractive summary that contains sentences from different evidence types.
ArgKP (Bar-Haim et al., 2020a)	controversial topics	28 topics, 378 key points, 7000 arguments	foundation dataset for KPA, includes arguments, key points and stance labels for each topic.
ArgKP21 (Friedman et al., 2021)	same as ArgKP	extend ArgKP by 3 topics as test set	Used in 2021 KPA shared task.
ArgCMV (Gurjar et al., 2025)	Reddit ChangeMyView	12K argument key point pairs	a key point extraction dataset created using real conversational threads.
Li et al. (2024a)	controversial topics from ArgKP	14K	A multitask dataset for end-to-end argument summarization and evaluation.

Table 3: Representative datasets for argument summarization.

Metric	Category	What it measures	Used in
ROUGE	textual overlap	N-gram overlap and longest common subsequence similarity between model output and human summaries.	DebateSum (Roush and Balaji, 2020), ConvoSumm (Fabbri et al., 2021a)
soft-precision, soft-recall, soft-F1 (Li et al., 2023)	semantic similarity	Soft-Precision measures how many predicted key points or argument units semantically match gold labels; Soft-Recall measures how many gold units are covered by predictions; Soft-F1 is the harmonic mean.	Altemeyer et al. (2025); Gurjar et al. (2025)
coverage score (Khosravani et al., 2024)	semantic similarity	The percentage of ground truth key points captured by the model predictions.	Altemeyer et al. (2025)
LLM-based coverage and redundancy (Altemeyer et al., 2025)	LLM as a judge	Prompting an LLM to evaluate the coverage and redundancy of the generated key points.	Altemeyer et al. (2025)

Table 4: Representative evaluation metrics for KPA.

or argument mining datasets, where argumentative structures are only used to improve the summary. Sukprapa et al. (2021) used the micro-text dataset (Peldszus, 2015) to test whether formal argumentation can improve text summarization. Furthermore, some datasets claim to support argument summarization, but in our view, they do not meet the strict definition of argument summarization we adopted in this paper. For instance, DebateSum (Roush and Balaji, 2020) is a dataset of debate evidence, where each instance contains an evidence, its extractive summary, and an associated argument. Although it contains an input text (evidence) and a corresponding summary, the evidence does not contain rich argumentative structures and the extractive summary does not explicitly preserve the argumentative structure. OpenDebateEvidence (Roush et al., 2024) suffers from the same limitation, although the dataset contains evidence alongside extractive and abstractive summaries, the evidence is not inherently argumentative and neither type of summary explicitly preserves argumentative structure.

Overall, argument summarization datasets are narrow and inconsistently designed, limiting reuse and hindering systematic comparison across works.

7 Evaluation Metrics

When evaluating argument summarization, KPA and ATS have different goals. In KPA, generated key points are compared with reference key points, with evaluation focusing on two complementary aspects: coverage, which measures how well the generated key points capture the meaning of the reference key points, and redundancy, which measures overlap among the generated key points. Generated key points can be evaluated with ROUGE either by concatenating generated key points and comparing them against concatenated reference key points (Li et al., 2023) or by aligning each generated key point with its best-matching reference key point and averaging the scores (van der Meer et al., 2024). However, purely textual overlap metrics fail to capture semantic equivalence between differently worded but equivalent key points, as highlighted by Li et al. (2023). To address this limitation, Li et al. (2023) introduced soft evaluation metrics, including soft-precision, soft-recall, and soft-F1. These metrics assess semantic simi-

larity between predicted and reference key points, typically using pre-trained scoring functions such as BLEURT (Sellam et al., 2020) or BARTScore (Yuan et al., 2021). Soft-precision measures the proportion of generated key points that semantically match references, soft-recall measures how much of the reference set is covered, and soft-F1 combines the two. Similarly, the coverage score proposed by Khosravani et al. (2024) computes the proportion of reference key points captured by model predictions using semantic similarity matching. More recently, Altemeyer et al. (2025) used LLMs as judges to assess coverage and redundancy directly. Overall, KPA evaluation has evolved from surface-level overlap to semantic similarity and, more recently, to LLM-based assessment, closely mirroring trends in generic NLP evaluation. Table 4 shows representative evaluation metrics for KPA.

For ATS, as the summary aims to preserve both semantic content and argumentative structure, the evaluation metrics need to measure both semantic and structural aspects. For example, Syed et al. (2020) measured model's summary by its adherence to the editorial's structure coverage of unique summary segments. Tang et al. (2026) measured the generated claim-reason structure of online conversations from multiple dimensions, including information quality, semantic quality, correctness, coverage and convincingness. Their methods combined automatic metrics, LLM and human evaluation. However, many ATS studies continue to rely primarily on textual overlap based metrics such as ROUGE and semantic similarity metrics such as BERTScore. For instance, although Elaraby et al. (2023) proposed generating argument-aware summaries structured around issues, reasons, and conclusions, evaluation was limited to surface-level semantic metrics, without explicitly measuring the preservation of argumentative structure. As a result, systems that differ substantially in how well they retain argumentative content may receive similar evaluation scores. Overall, the evaluation of ATS outputs, especially from the structural aspect, remains challenging.

8 Argument Summarization in the Broader NLP Landscape

Argument summarization lies at the intersection of several major NLP research areas, most notably generic text summarization. This section compares generic and argument summarization by examining datasets, modeling approaches, and evaluation practices, and highlighting the resulting gaps and challenges specific to argument summarization.

8.1 Generic Text Summarization

Generic text summarization is a long-standing task that has progressed alongside major advances in NLP, evolving from statistical and fully supervised machine learning methods (Luhn, 1958; Nallapati et al., 2017) to PLMs (Liu and Lapata, 2019) and LLMs (Wang et al., 2023; Fang et al., 2025). This development has been driven by the rich datasets for text summarization, including short summarization datasets in news (Hermann et al., 2015) and other domains (Gliwa et al., 2019), long document (Sharma et al., 2019) and multi-document (Over et al., 2007) summarization datasets, and application-specific datasets such as meeting summarization (Zhong et al., 2021) or aspect-based summarization (Bahrainian et al., 2022). The success of LLMs has advanced text summarization research, enabling a wide range of methods such as various prompting methods (Ravaut et al., 2023; Wang et al., 2023), multi-agent methods (Fang et al., 2025), alignment learning (Liu et al., 2023; Song et al., 2025), and distillation (Jiang et al., 2024), with applications covering domains such as news (Wang et al., 2023), medical domain (Tang et al., 2023), or scientific domain (Li et al., 2024b).

Evaluation has been widely studied with overlap-based methods (ROUGE (Lin, 2004)), PLM-based methods (similarity based methods: BERTScore (Zhang et al., 2019), BARTScore (Yuan et al., 2021)), NLI-based methods (Tang et al., 2024b), and LLM-based methods (Song et al., 2024). Evaluation has covered multiple dimensions: faithfulness (Bhandari et al., 2020; Pagnoni et al., 2021; Laban et al., 2023), coherence and relevance (Fabbri et al., 2021b), completeness and conciseness (Lee et al., 2024). The granularity has shifted from coarse-grained, summary level evaluation (Fabbri et al., 2021b; Laban et al., 2023) to fine-grained, sentence level evaluation (Lee et al., 2024).

8.2 Argument Summarization: Current Limitations in Practice

Compared to generic text summarization, argument summarization has substantially fewer datasets, methods, and evaluation resources. For KPA, only three datasets (ArgKP, ArgKP21, ArgCMV) exist, all focused on online discussions (Section 6). This narrow dataset coverage restricts both domain

generalization and methodological diversity. LLM-based KPA is sparse (Tang et al., 2024a; Altemeyer et al., 2025), the latter being the only work that systematically studies LLMs for KPA. However, their focus is on prompting-based, single-LLM approaches. Compared to the multi-dimension, multi-granularity and rich methods evaluation for generic text summarization, the evaluation for KPA also lags behind, with only a few studies addressing coverage and redundancy (Li et al., 2023; Khosravani et al., 2024), and existing LLM-based evaluations lacking in-depth analyses (Altemeyer et al., 2025).

ATS faces similar limitations. Many datasets (e.g., WebisEditorialSum-2020 (Syed et al., 2020), ConvoSumm (Fabbri et al., 2021a)) are not specifically designed for argument summarization, but instead use argumentative structures for dataset creation, while others (e.g., DebateSum (Roush and Balaji, 2020), OpenDebateEvidence (Roush et al., 2024)) can only be used for generic text summarization. To our knowledge, only Li et al. (2024a) contains a subset for ATS with claim-evidence pairs as inputs and abstractive summaries that preserve the argumentative structure as outputs. Consequently, most ATS works only use argumentative structure for generic text summarization, while methods that can explicitly produce summaries preserving the argumentative structure of the input remain scarce.

8.3 Unique Challenges in Argument Summarization

Argument summarization faces several challenges that distinguish it from generic text summarization.

Relational argumentative meaning. Unlike generic summarization, argument summarization must preserve *relations* between propositions (e.g., support/attack), not just salient content. Summaries that retain key claims while omitting these relations may yield fluent but distorted argumentation.

Compression versus structural fidelity. Argumentative structures often encode richer implicit structure than is explicitly expressed in text. Compressing them into concise summaries risks collapsing argumentative roles while insufficient compression undermines the goal of summarization.

Heterogeneous inputs and outputs. Prior work spans single long texts, multi-turn discussions, and sets of short arguments, with outputs ranging from free-form summaries to key points or structured representations. This heterogeneity hinders model reuse and consistent evaluation across settings.

Evaluation limitations. Evaluating argument summarization is difficult as it requires assessing both semantic adequacy and if the input's *argumentative meaning* is preserved. Metrics like ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2019) (see Section 7) capture surface-level or semantic similarity, but are largely insensitive to the preservation of argumentative relations. LLM-based evaluations adopted in KPA (Altemeyer et al., 2025) assess coverage and redundancy, but are difficult to generalize to other tasks in argument summarization.

9 Future Directions

Dataset development and standardization. Current progress in argument summarization is limited by dataset scarcity and fragmentation. Unlike generic text summarization, which benefits from large-scale standardized benchmarks across domains and summary types, argument summarization datasets are domain-specific and heterogeneous in both input structure and output format. For ATS, most datasets either adapt generic summarization corpora with argumentative signals as auxiliary information or rely on narrow, domain-specific annotations. For KPA, existing benchmarks are limited to short, independent arguments and a small set of topics, limiting generality. Future dataset development should focus on clearly defined task formulations with explicit alignment between inputs, outputs, and argumentative annotations. Datasets should provide multi-granularity argumentative annotations, cover diverse input structures beyond isolated arguments, and enable systematic comparison across ATS and KPA settings. Establishing shared benchmarks and evaluation splits would improve reproducibility and comparability across studies.

Structure-aware and multi-dimensional evaluation. Evaluation remains a key challenge. Future metrics should move beyond semantic similarity and explicitly assess whether argumentative structures are preserved. Promising directions include relation-aware matching metrics, graph-based comparisons, and hybrid human-automatic evaluation. Multi-dimensional and multi-granularity frameworks should jointly assess semantic quality, structural faithfulness, coverage, and redundancy.

Principled use of large language models. While generic text summarization has explored prompting, multi-agent frameworks, alignment learning, and distillation, LLM-based argument summarization has mainly relied on prompting and pipeline-based integration. Future work should adapt these

methods to argument-aware settings, e.g. through argument-aware prompting strategies, multi-agent summarization frameworks that reason over argumentative roles and viewpoints, and alignment or distillation approaches that explicitly preserve argumentative relations. Ensuring that LLM-generated summaries remain faithful to the input's argumentative structure is a key challenge for ATS and KPA.

10 Conclusion

Argument summarization is an important but conceptually fragmented research area. This survey provided a unified definition and introduced a taxonomy along input structure, output representation, and the role of argumentative structure. Using this framework, we analyzed the field's evolution and reviewed existing datasets, methods, and evaluation practices, identifying systematic misalignments that hinder comparability and progress. We highlighted the need for better-aligned datasets, more structure-aware evaluation, and principled use of LLMs to ensure argumentative faithfulness. We hope this survey offers a clearer foundation for future research and supports more coherent advances in argument summarization.

11 Limitations

This survey has several limitations. First, although we aim for comprehensive coverage, the scope of argument summarization is broad and rapidly evolving, particularly with the increasing use of large language models. As a result, some recent or domain-specific works may not be fully captured, and our analysis reflects the state of the literature at the time of writing.

Second, our taxonomy and unified task definition necessarily abstract over diverse task formulations and annotation schemes. While this abstraction is intended to clarify commonalities and sources of fragmentation, it may oversimplify certain task- or domain-specific nuances, especially in highly specialized applications such as legal or scientific argumentation.

Third, our analysis is based on reported datasets, methods, and evaluation practices in the literature, rather than on new empirical experiments. Consequently, some of our conclusions about misalignment and evaluation limitations are supported by comparative reasoning and reported findings, rather than direct experimental validation.

Finally, while we discuss future directions such as dataset standardization, structure-aware evaluation, and principled use of large language models, we do not provide concrete implementations or benchmarks for these directions. Developing and validating such resources remains an open challenge for future work.

References

Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019. Modeling frames in argumentation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2922–2932, Hong Kong, China. Association for Computational Linguistics.

S Alshaina, Ansamma John, and Aneesh G Nath. 2017. Multi-document abstractive summarization based on predicate argument structure. In *2017 IEEE international conference on signal processing, informatics, communication and energy systems (SPICES)*, pages 1–6. IEEE.

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

Moritz Altemeyer, Steffen Eger, Johannes Daxenberger, Yanran Chen, Tim Altendorf, Philipp Cimiano, and Benjamin Schiller. 2025. Argument summarization and its evaluation in the era of large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35478–35499.

Seyed Ali Bahrainian, Sheridan Feucht, and Carsten Eickhoff. 2022. NEWTS: A corpus for news topic-focused summarization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 493–503, Dublin, Ireland. Association for Computational Linguistics.

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*, pages 4029–4039, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. 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.

Charlie Egan, Advaith Siddharthan, and Adam Wyner. 2016. Summarising the points made in online political debates. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 134–143, Berlin, Germany. Association for Computational Linguistics.

Mohamed Elaraby and Diane Litman. 2022. ArgLegalSumm: Improving abstractive summarization of legal documents with argument mining. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6187–6194, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Mohamed Elaraby, Yang Zhong, and Diane Litman. 2023. Towards argument-aware abstractive summarization of long legal opinions with summary reranking. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7601–7612, Toronto, Canada. Association for Computational Linguistics.

Alexander Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021a. ConvoSumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. 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 6866–6880, Online. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021b. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Jiangnan Fang, Cheng-Tse Liu, Jieun Kim, Yash Bhedaru, Ethan Liu, Nikhil Singh, Nedim Lipka, Puneet Mathur, Nesreen K. Ahmed, Franck Dernoncourt, Ryan Rossi, and Hanieh Deilamsalehy. 2025. Multi-LLM text summarization. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 352–362, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Roni Friedman, Lena Dankin, Yufang Hou, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2021. Overview of the 2021 key point analysis shared task. In *Proceedings of the 8th Workshop on Argument Mining*, pages 154–164, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Omkar Gurjar, Agam Goyal, and Eshwar Chandrasekharan. 2025. ArgCMV: An argument summarization benchmark for the LLM-era. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21870–21883, Suzhou, China. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Pengcheng Jiang, Cao Xiao, Zifeng Wang, Parminder Bhatia, Jimeng Sun, and Jiawei Han. 2024. TriSum: Learning summarization ability from large language models with structured rationale. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2805–2819, Mexico City, Mexico. Association for Computational Linguistics.

Manav Kapadnis, Sohan Patnaik, Siba Panigrahi, Varun Madhavan, and Abhilash Nandy. 2021. Team enigma at ArgMining-EMNLP 2021: Leveraging pre-trained language models for key point matching. In *Proceedings of the 8th Workshop on Argument Mining*, pages 200–205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mohammad Khosravani, Chenyang Huang, and Amine Trabelsi. 2024. Enhancing argument summarization: Prioritizing exhaustiveness in key point generation and introducing an automatic coverage evaluation metric. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8212–8224, Mexico City, Mexico. Association for Computational Linguistics.

Philippe Laban, Wojciech Kryscinski, Divyansh Agarwal, Alexander Fabbri, Caiming Xiong, Shafiq Joty, and Chien-Sheng Wu. 2023. SummEdits: Measuring LLM ability at factual reasoning through the lens

of summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9662–9676, Singapore. Association for Computational Linguistics.

Yuho Lee, Taewon Yun, Jason Cai, Hang Su, and Hwanjun Song. 2024. UniSumEval: Towards unified, fine-grained, multi-dimensional summarization evaluation for LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3941–3960, Miami, Florida, USA. Association for Computational Linguistics.

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.

Hao Li, Yizheng Sun, Viktor Schlegel, Kailai Yang, Riza Batista-Navarro, and Goran Nenadic. 2025. Arg-llada: Argument summarization via large language diffusion models and sufficiency-aware refinement. *arXiv preprint arXiv:2507.19081*.

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

Zekai Li, Yanxia Qin, Qian Liu, and Min-Yen Kan. 2024b. Isqa: Informative factuality feedback for scientific summarization. *arXiv preprint arXiv:2404.13246*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

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

Yixin Liu, Budhaditya Deb, Milagro Teruel, Aaron Halfaker, Dragomir Radev, and Ahmed Hassan Awadallah. 2023. On improving summarization factual consistency from natural language feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15144–15161, Toronto, Canada. Association for Computational Linguistics.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

Amita Misra, Brian Ecker, and Marilyn Walker. 2016. Measuring the similarity of sentential arguments in dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 276–287, Los Angeles. Association for Computational Linguistics.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Paul Over, Hoa Dang, and Donna Harman. 2007. Duc in context. *Information Processing & Management*, 43(6):1506–1520.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Andreas Peldszus. 2015. An annotated corpus of argumentative microtexts.

Viet Hoang Phan, Tien Long Nguyen, Duc Long Nguyen, and Ngoc 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, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mathieu Ravaut, Hailin Chen, Ruochen Zhao, Chengwei Qin, Shafiq Joty, and Nancy Chen. 2023. Promptsum: Parameter-efficient controllable abstractive summarization. *arXiv preprint arXiv:2308.03117*.

Jan Heinrich Reimer, Thi Kim Hanh Luu, Max Henze, and Yamen Ajjour. 2021. Modern talking in key point analysis: Key point matching using pretrained encoders. In *Proceedings of the 8th Workshop on Argument Mining*, pages 175–183, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578, Florence, Italy. Association for Computational Linguistics.

Allen Roush and Arvind Balaji. 2020. DebateSum: A large-scale argument mining and summarization dataset. In *Proceedings of the 7th Workshop on Argument Mining*, pages 1–7, Online. Association for Computational Linguistics.

Allen Roush, Yusuf Shabazz, Arvind Balaji, Peter Zhang, Stefano Mezza, Markus Zhang, Sanjay Basu, Sriram Vishwanath, Mehdi Fatemi, and Ravid S Ziv. 2024. Opendebateevidence: A massive-scale argument mining and summarization dataset. *Advances in Neural Information Processing Systems*, 37:34270–34293.

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, Online. Association for Computational Linguistics.

Huawei Shan and Dongyuan Lu. 2025. Debate divides: Argument relation-based contrastive opinion summarization via multi-task learning for online discussions. *Neurocomputing*, 638:130124.

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024. FineSurE: Fine-grained summarization evaluation using LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 906–922, Bangkok, Thailand. Association for Computational Linguistics.

Hwanjun Song, Taewon Yun, Yuho Lee, Jihwan Oh, Gihun Lee, Jason Cai, and Hang Su. 2025. Learning to summarize from llm-generated feedback. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 835–857.

Isada Sukprapa, Nguyen Duy Hung, and Thepchai Supnithi. 2021. Text summarization using formal argumentation. In *2021 16th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, pages 1–6. IEEE.

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

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

An Quang Tang, Xiuzhen Zhang, Minh Ngoc Dinh, and Erik Cambria. 2024a. Prompted aspect key point analysis for quantitative review summarization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10691–10708, Bangkok, Thailand. Association for Computational Linguistics.

An Quang Tang, Xiuzhen Zhang, Minh Ngoc Dinh, and Zhuang Li. 2026. Arqusumm: Argument-aware quantitative summarization of online conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 33205–33213.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024b. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, and 1 others. 2023. Evaluating large language models on medical evidence summarization. *NPJ digital medicine*, 6(1):158.

Michiel van der Meer, Piek Vossen, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2024. An empirical analysis of diversity in argument summarization. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2028–2045, St. Julian's, Malta. Association for Computational Linguistics.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8640–8665, Toronto, Canada. Association for Computational Linguistics.

Huihui Xu, Jaromir Savelka, and Kevin D Ashley. 2021. Toward summarizing case decisions via extracting argument issues, reasons, and conclusions. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 250–254.

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. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

Wendi Zhou, Ameer Saadat-Yazdi, and Nadin Kökciyan. 2025. Aspect-based opinion summarization with argumentation schemes. In *Proceedings of the 12th Argument mining Workshop*, pages 116–125, Vienna, Austria. Association for Computational Linguistics.

Timon Ziegenbein, Shahbaz Syed, Martin Potthast, and Henning Wachsmuth. 2024. Objective argument summarization in search. In *Conference on Advances in Robust Argumentation Machines*, pages 335–351. Springer.