Responsible AI Considerations in Automatic Text Summarization Research

Yu Lu Liu

School of Computer Science

McGill University, Montreal

August 1, 2024

A thesis submitted to McGill University in partial fulfillment of the requirements

of the degree of Master of Science

Acknowledgements

I would like to thank Prof. Jackie Chi Kit Cheung for teaching me how to think and work as a researcher, how to communicate and collaborate, and how to not despair when a project stagnates. I'm very grateful for his guidance and support throughout my journey. Furthermore, I extend my gratitude to Dr. Alexandra Olteanu and Dr. Su Lin Blodgett for their mentorship, both within and beyond this project. Their work, their experiences, and their perspectives have greatly inspired me to continue my journey as a researcher. I'm also grateful to have my research potential recognized and supported by Fonds de Recherche du Québec Nature et Technologies.

I'm grateful to be surrounded by wonderful labmates and friends. I would like to especially thank Cesare Spinoso-Di Piano and Ziling Cheng for all the lively conversations we shared, which brightened up even the dullest days at the lab. I also need to thank my longtime best friends Lucy Truong and Frédéric Duong for always being there for me, for dragging me out of the house for my own good, and for ensuring I never forget about my embarrassing moments from high school.

Finally, I would not be where I am today without my parents, who have sacrificed so much and worked so hard to build a new life here in Canada. I am forever grateful for their love, their unwavering faith in me, and of course their excellent sense of humour, all of which making home truly the happiest place on earth. This thesis is thus dedicated to 爸爸妈妈.

Abstract

The task of automatic text summarization (ATS) involves automatically producing a shorter version of an input text while preserving important information. As summaries allow readers to quickly grasp the main ideas of textual data, ATS systems have a wide range of applications, ranging from producing news article headlines to summarizing customer reviews. System-generated summaries, however, could be incorrect, biased and even harmful — e.g., a news headline incorrectly representing the events reported in the source article, thereby misinforming the public. Despite increasing interest in the responsible development and use of artificial intelligence technologies, there is — for the task of ATS — a lack of understanding as to how prevalent such issues are, or when and why these issues are likely to arise. To contribute to our understanding of responsible artificial intelligence (RAI) issues in ATS, we examine research and reporting practices in the current ATS literature by conducting a multi-round content analysis of 333 contemporary ATS papers. We focus on how, which, and when responsible AI issues are covered, which relevant stakeholders are considered, and mismatches between stated and realized research goals. We also discuss current evaluation practices and consider how authors discuss the limitations of both prior work and their own work. Overall, we find that relatively few papers engage meaningfully with RAI through their research and reporting practice, which limits their consideration of potential downstream adverse impacts or other RAI issues. Based on our findings, we make recommendations on concrete practices and research directions.

Abrégé

Le résumé automatique de texte est un processus qui consiste à produire automatiquement une version plus courte d'un texte tout en conservant ses informations pertinentes. Les résumeurs automatiques peuvent être utilisés dans divers contextes, comme pour produire automatiquement des titres d'actualité, ou pour résumer des avis clients sur des plateformes e-commerce. Cependant, les résumés produits par ces technologies pourraient être incorrects, biaisés et même nuisibles : par exemple, un titre d'actualité représentant de manière incorrecte les événements rapportés dans un article de journal peut ainsi désinformer le public. Malgré un intérêt croissant pour le développement et l'utilisation responsable des technologies d'intelligence artificielle (IA responsable), nous comprenons mal l'ampleur de ces problèmes, ou encore quand et pourquoi ces problèmes sont susceptibles de survenir pour les résumeurs automatiques. Pour contribuer à une meilleure compréhension des problèmes d'IA responsable dans le cadre du résumé automatique, nous examinons les pratiques de la recherche dans la littérature en effectuant une analyse de contenu sur 333 ouvrages récemment publiés sur le sujet de résumé automatique. Nous nous intéressons aux questions sur comment, quelles et quand l'IA responsable est abordé, sur les parties prenantes (*stakeholders*) pertinentes, et sur les écarts entre les objectifs de recherche déclarés et réalisés. Nous étudions également les pratiques d'évaluation utilisées dans ces ouvrages et examinons la manière dont les auteurs discutent des limites des études antérieures et de leur propre étude. Dans l'ensemble, nous constatons que relativement peu d'ouvrages s'occupent significativement du sujet de l'IA responsable à travers leurs pratiques de recherche. Cela limite leur prise en compte des impacts négatifs possibles ou d'autres problèmes associée à l'IA responsable. En tenant compte de ces résultats,

nous fournissons des recommandations sur des pratiques concrètes et des directions futures de recherche.

Author Contributions

The content of this thesis is adapted from the publication at the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP):

Yu Lu Liu, Meng Cao, Su Lin Blodgett, Jackie Chi Kit Cheung, Alexandra Olteanu, Adam Trischler. 2023. Responsible AI Considerations in Text Summarization Research: A Review of Current Practices. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6246–6261, Singapore. Association for Computational Linguistics.

This project is led by Yu Lu Liu, who coordinated the paper annotation process (including training and management of hired annotators) and conducted analysis of resulting annotations. This project is first proposed by Su Lin Blodgett, Jackie Chi Kit Cheung, Alexandra Olteanu, and Adam Trischler. All team members contributed to annotating examined papers, to discussions, and to paper writing.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Automatic text summarization (ATS) is an important natural language processing (NLP) task where the goal is to produce a shorter version of an input text while preserving its main ideas. The task has a wide range of possible applications. Some examples include summarizing news articles (e.g., to produce news headlines), medical reports, research papers, meeting transcripts, and emails (e.g., to produce email commitment reminders).

As the capabilities of ATS systems increase, especially with the emergence of large language models, they have seen increasing use despite the known risks of generating incorrect, biased, or otherwise harmful summaries. Generated summaries might, for instance, misgender the people they describe; give rise to libelous representations by failing to appropriately qualify claims (e.g., *"The suspected murderer ran away"* v.s. *"The murderer ran away"*); mislead users by giving rise to inferences that are ambiguous or unsupported by the source text; represent contested topics unfairly (e.g., only summarizing arguments presented by one side of a debate); or be susceptible to adversarial perturbations in the source text. Such cases of failure could even be indicative of harmful biases that the system has, e.g., if utterances by female employees are systematically overlooked by a meeting ATS. The responsible development and use of ATS systems thus require understanding these risks: how likely they are to occur, what impacts they have on stakeholders, which system development practice lead to biased systems, etc.

There is a growing body of research examining responsible artificial intelligence (RAI) concerns in the field of NLP, and in AI in general (Boyarskaya et al., 2020; Nanayakkara et al., 2021; Hardmeier et al., 2021). There have also been growing efforts to incorporate, in AI and NLP research practice, reflections about ethical considerations, adverse impacts, and other RAI issues that NLP and AI research—and related applications—can exacerbate (Benotti and Blackburn, 2022; Ashurst et al., 2022; Nanayakkara et al., 2021; Boyarskaya et al., 2020). As an example, Empirical Methods in Natural Language Processing (EMNLP)[1] a leading conference in the field of NLP, made a series of changes related to RAI in recent years: In 2020, the conference introduced an ethics policy; in 2021, authors were allowed authors extra space for a broader impact statement or other discussion of ethics; finally, EMNLP 2022 required all submitted papers to contain a dedicated section on discussions of limitations. These new guidelines and requirements reflect the increasing importance of RAI in the NLP research community.

1.1 Thesis Outline

Despite efforts to encourage reflections around RAI-related issues, and despite the array of risks associated with ATS system failures, little work has comprehensively examined RAI concerns arising from ATS systems. In this thesis, we make a step forward in addressing this gap by investigating research and reporting practices related to how, when, and which RAI issues are covered in the contemporary ATS literature.

To examine research and reporting practices, we developed a set of annotation guidelines targeting aspects relevant to RAI through a multi-round annotation process. Following these guidelines, we conducted a detailed, systematic review of 333 ATS papers published between 2020 and 2022 in the ACL Anthology, a repository of papers maintained by the Association for Computational Linguistics (ACL) that currently hosts over 90K papers on the study of computational linguistics and NLP. Specifically, we examine how authors discuss limitations of both prior work

[1] EMNLP 2020, 2021, 2022

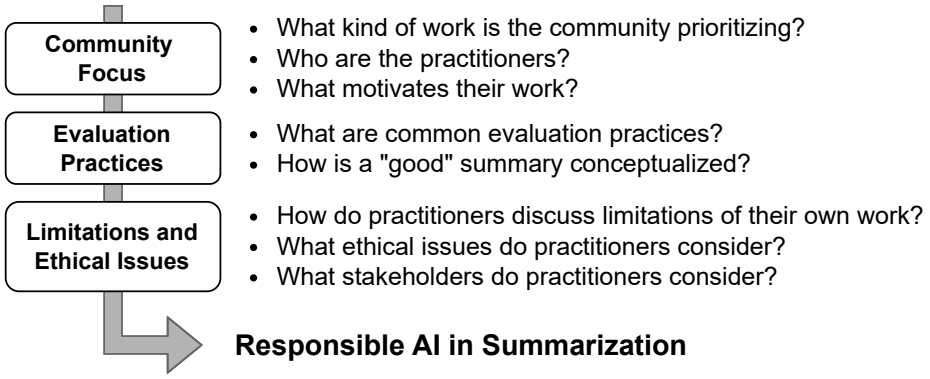

Figure 1.1: Overview of questions we examine when analyzing practices related to how the contemporary text summarization literature engages with RAI issues.

and their own work, which RAI issues they consider, the relevant stakeholders they imagine or serve, as well as how stated and realized research goals might often differ.

From our survey, we find that i) relatively few papers engage with possible stakeholders or contexts of use; ii) common evaluation practices may not provide meaningful insights about systems' true performance; iii) papers rarely engage with limitations and ethical concerns. We further discuss how this limits the space of RAI issues of which the community may be aware, as well as the community's capacity to speculate effectively on potential issues. Finally, we make recommendations for how the community can improve ATS research practices to be more responsible.

1.2 Statement of Contributions

In summary, this thesis makes the following contributions:

- We build a set of annotation guidelines that allow us to track research and reporting practices in ATS research community, from a RAI perspective.
- Following the above guidelines, we conduct a systematic review of over 300 ATS papers. The resulting annotations form a valuable dataset that we make available upon request[2].

[2]https://forms.gle/6J4wpp8daNJYjWMPA

- We extract findings from our review, providing insights about research and reporting practices in ATS research community (illustrated in Figure 1.1). We further discuss their implications in the context of RAI and propose recommendations.

This thesis helps foreground our choices as a community—about how we write, how we frame problems, how we consider social context, and how we broadly think about RAI issues—and to make these choices explicit (rather than implicit) so that we may better understand their implications. Since the NLP community has only recently started to prioritize these issues, taking an early snapshot of emerging practices can provide insight into why the community might be struggling with considering limitations of its work, ethical considerations, adverse impacts, and other RAI-related issues.

1.3 Organization of the Thesis

This thesis is composed of 6 chapters. The present chapter, Chapter 1, is an introduction to the thesis. Chapter 2 provides background on various topics relevant to this thesis. It contains an overview on the task of ATS: common approaches of ATS systems, datasets, and evaluation methods. It also contains an overview of prior work on RAI, specifically in the field of NLP, and then in the context of ATS. Chapter 3 describes the methodology of our survey: our paper collection and paper annotation process. This includes detailing our annotation scheme and highlighting the relevance of each annotated aspect to RAI. Chapter 4 presents findings from our survey, and in light of these findings, we present in Chapter 5. Finally, Chapter 6 summarizes the findings of this thesis, describes limitations of our work, and offer future directions.

Chapter 2

Background

In this chapter, we provides background on the task of automatic text summarization (ATS) and prior work on responsible AI (RAI) that addresses ATS. To further provide context for our work, we present an overview of RAI in general, and in the field of natural language processing (NLP). More specifically, in Section 2.1, we gives a brief survey of ATS: i) various approaches employed by ATS systems, ii) various domains of application and commonly used datasets, and iii) evaluation approaches for ATS systems. In Section 2.2, we provide background on RAI, specially prior work in the field of NLP in Section 2.2.1. This body of work informs our methodology, the implications of our findings, as well as the recommendations we provide to the community in light of our findings. To further contextualize our work, we summarize prior work on RAI addressing ATS in Section 2.2.2.

2.1 Automatic Text Summarization

ATS is a longstanding NLP research topic. Given one or more source texts, the main objective of text summarization systems is to produce a shorter text that convey key information of the input.

2.1.1 Approaches

Methods for producing summaries can be categorized into two main types: extractive and abstractive (Hahn and Mani, 2000).

Extractive Approaches

Extractive ATS builds the output summary by extracting passages (e.g., sentences) from the original document. These approaches often involve modeling the "importance"[1] of passages and concatenating the most important passages to form the output summary. The main concern with extractive summaries is their lack of coherence, as concatenated passages might not flow naturally or logically from one to the next. Furthermore, extractive approaches are limited in the type of reasoning or transformation they can perform on the original document, which may be necessary in some contexts. For example, summarizing a meeting transcript may require converting passages in first-person into third-person (e.g., *"Alice: I agree!"* to *"Alice agrees."*), which is impossible in extractive ATS.

Abstractive Approaches

Abstractive ATS generates the output summary: it requires producing new text that summarizes important ideas from the original document. Most abstractive ATS systems harness the capabilities of language models. For example, an abstractive ATS system could be obtained by further training (i.e. "fine-tuning") BART (Lewis et al., 2020). BART uses a bidirectional encoder and an autoregressive decoder. During pre-training, it learns to reconstruct the original text from corrupted texts it receives as inputs. Then, to adapt it to the task of summarization, the model is further trained on datasets of source-summary pairs (e.g., news articles paired with their headlines), where it learns to produce summaries from source documents it receives as inputs. The objective in this second training phase is somewhat similar to the model's pre-training objective, so the model would hope-

[1] Here, the term "importance" is used as an umbrella term for any desired attributes of the information the output summaries aim to contain. The exact attributes is dependant on the context. For example, for medical reports, patients might want the summaries to cover information on diagnosis, whereas medical practitioners might find information on symptoms and medical test results more important.

fully make use of capabilities it learned during pre-training (e.g., what makes a text coherent and grammatically correct) to perform the task of summarization.

A growing emphasis has been placed on ensuring that generated summaries are consistent with the source text, as abstractive systems risk generating so-called "hallucinations," i.e., text that distorts or is unsupported by the source text (Cao et al., 2020; Dong et al., 2020; Falke et al., 2019; Kryscinski et al., 2020; Kumar and Cheung, 2019). Related concerns about factuality, accuracy, and coherency have all been bundled under hallucinations, obscuring what issues the authors are after and the range of harms or adverse impacts they can bring about.

2.1.2 Datasets and Domains of Application

ATS has a wide range of applications, covering many domains. Here, we offer a brief, *non-exhaustive* overview of the domains and some commonly-used English-language datasets:

- **News** summarization often involve producing a headline or a highlights section for a given news article. Commonly used datasets include CNN-Dailymail (Hermann et al., 2015), Gigaword (Graff et al., 2003) and XSum (Narayan et al., 2018).

- **Dialogue** or conversation summarization include datasets of online conversations such as SAMSum (Gliwa et al., 2019), a corpus of messenger-like online dialogues, and ConvoSumm (Fabbri et al., 2021a), which covers news comments, discussion forums, and email threads. For summarizing professional meeting transcripts, there are corpora such as AMI (McCowan et al., 2005), ICSI (ICS, 2003), and Elitr (Nedoluzhko et al., 2022).

- **Opinion** summarization often take the involve summarizing multiple documents (e.g., customer reviews, social media posts expressing opinions). Opinion summarization corpora include Opinosis (Ganesan et al., 2010), OPOSUM (Angelidis and Lapata, 2018), and corpora extracted from platforms such as Yelp, Amazon, and Twitter.

- **Medical** summarization includes summarizing medical literature. MSˆ2, for example, is a dataset supporting multi-document summarization of medical studies. Medical summarization also involve medical records and reports, e.g., MEDIQA challenge for summarization

of radiology reports (Ben Abacha et al., 2019; Johnson et al., 2019). Due to patient privacy concerns, some datasets may be not released publicly (e.g., MedicalSum (Michalopoulos et al., 2022)).

Different domains of application may involve different risks. Our work examines how the research community engage with domains of application through their (stated) intended domains and their choice of training and evaluation corpora.

2.1.3 Evaluation

ATS systems are often evaluated using automatic reference-based metrics — metrics which automatically output quality scores for system-generated summaries by comparing them with reference summaries. They often rely on computing lexical overlap, a notable example being **ROUGE** (Lin, 2004), which stood for Recall-Oriented Understudy for Gisting Evaluation. It is one of the first and mostly widely adopted metrics for ATS, and it has many variant depending on the type of unit being considered in the computation of lexical overlap. ROUGE-N, for example, considers n-grams (i.e., sequences of n adjacent words): the ROUGE-N score between an output summary and a reference summary is the number of matching n-grams between the two texts, divided by the total number of n-grams in the reference summary. We show in Figure 2.1 the computation of ROUGE-2 through toy examples.

The figure also illustrates when automatic metrics based on lexical overlap could fail to capture the semantic similarity between output summaries and reference summaries: despite describing a different event, summary A (*"the lazy dog jumps over the quick brown fox"*) obtained a higher ROUGE-2 score compared to summary B (*"the fox jumps over the dog"*) because its bigrams matched more with the reference text(*"the quick brown fox jumps over the lazy dog"*). Indeed, these evaluation methods often do not correlate well with human judgments, especially on other domains than news. For instance, Liu and Liu (2008) found that ROUGE scores have a low correlation with human-annotated scores for extractive meeting summarization, and suggested improvements by taking into account several unique characteristics of the use context such as disfluencies (e.g., incomplete sentences) and speaker information. Cohan and Goharian (2016) arrived at sim-

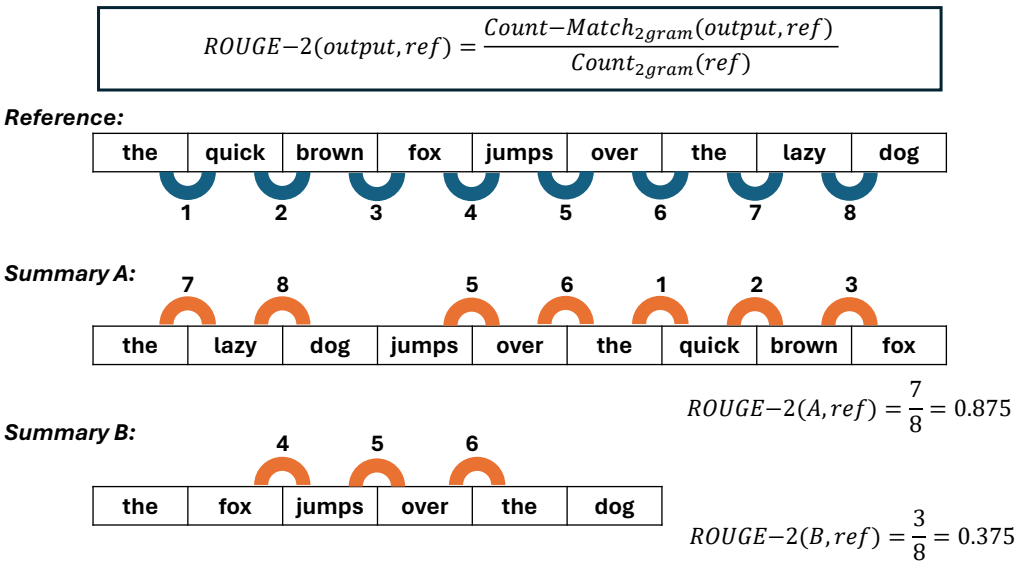

Figure 2.1: Toy examples to illustrate the score computation for ROUGE-2. Bigrams from the reference text are indicated by numbered blue half-arcs. For each output summary, bigrams that also occur in the reference text are indicated by orange half-arcs, with the number referring to the exact bigram from the reference text that is matched (e.g., "the lazy" from summary A matches with the 7th bigram in the reference text).

ilar conclusions for the use of ROUGE in the context of scientific document summarization. The shortcomings with existing metrics have spurred research into developing automatic metrics with higher human correlations (Fabbri et al., 2021b). Nevertheless, automatic metrics can obfuscate when systems may or may not work; e.g., are the summaries more likely to leave out a certain type of content? Do they work equally well for content by different speakers?

Beyond automatic metrics, collecting human judgements is a commonly used evaluation method, often involving asking human annotators to rate system-generated summaries in terms of their overall quality or of intrinsic quality criteria such as coherence, relevance, etc (Gkatzia and Mahamood, 2015). More extrinsic criteria related to how summaries are used downstream applications (i.e., usability) are rarely considered (Zhou et al., 2022).

Our work examines how evaluation practices might limit the space of RAI issues the community considers by possibly obfuscating some issues – e.g., failure to evaluate a criteria may conceal

issues related to said criteria – and foregrounding others through, for instance, extensive evaluation of criteria that reveal related issues.

2.2 Responsible AI

The responsible development of AI systems involve various ethical principles that a community values. For example, the Montreal Declaration for a Responsible Development of Artificial Intelligence[2] highlight 10 principles: the protection of privacy, democratic participation, equity, sustainability, and others. Different communities may value different principles. By comparing over 30 prominent documents on AI principles (spanning across corporations, academic institutions, governments, etc.), Fjeld et al. (2020) found common values, as well as differences in interpretation. For instance, some documents treat the principle of "transparency" as binary (i.e., an AI system is either transparent or not), whereas others consider different levels to the principle. As RAI envelops these diverse ethical principles that may vary from community to community, it is impossible to impose a strict definition for this term. In the following sections, we describe prior work related to RAI (i.e., related to principles such as fairness) that are relevant to this thesis.

2.2.1 Responsible AI in NLP

A large body of prior work discussed and studied RAI-related concerns in NLP, more specifically about language models which support most contemporary NLP systems (Weidinger et al., 2022; Bender, 2019, i.a.). These models are known to learn harmful social biases (Bolukbasi et al., 2016; Caliskan et al., 2017; Zhao et al., 2019, i.a.), e.g., model representations associating European American names with pleasantness while associating African American names with unpleasantness (Caliskan et al., 2017). Blodgett et al. (2020b) connect uses of models that encoded such biases to allocational harms (e.g., NLP systems that sort job applicants' resumes allocating less career opportunities to people with African American names) and representational harms (e.g., stereotypes in system outputs). These concerns are related to principles like fairness, non-

[2] https://montrealdeclaration-responsibleai.com/)

discrimination, equity and diversity. Privacy-related concerns have also been raised, with Carlini et al. (2021) showing that it is possible to "trick" (through adversarial prompt attacks) a large language model into reproducing personally identifiable information.

The development of NLP systems itself can have negative societal impacts. The high environmental cost of training large language models — the training procedure of a Transformer model is estimated to emit over 250 tons of CO_2 (Strubell et al., 2019)— is directly connected to principles about sustainability. The ethical aspect of the data collection process is greatly relevant, as many NLP systems rely on human annotators to provide training data and output quality assessment. Rowe; Tan and Cabato; Ludec et al. (2023, i.a.) report and discuss ethical concerns about the outsourcing of data annotation labour into the Global South.

Meta-Analyses in NLP

Meta-analyses in NLP — work that analyzes research and reporting practices in NLP — help reveal weaknesses in current practices. Blodgett et al. (2020a) explore how NLP papers describe "bias," finding that definitions are often vague, vary widely, and may not be well-matched to accompanying technical approaches, while Benotti and Blackburn (2022) examine ethical considerations sections in ACL 2021 papers, finding that relatively few (∼15%) include such sections, and that some of these (∼20%) do not meaningfully address either benefits or harms of the research. Blodgett et al. (2021) examine how four benchmark datasets conceptualize and operationalize stereotyping, while Devinney et al. (2022) analyze papers on gender bias in NLP to uncover how gender is theorized, finding that theorizations rarely are made explicit or engage with gender theories beyond NLP. Via interviews and a survey, Zhou et al. (2022) examine practitioners' assumptions and practices when evaluating natural language generation systems. Elsewhere, work has analyzed evaluation practices in natural language generation (Gkatzia and Mahamood, 2015; van der Lee et al., 2019; Howcroft et al., 2020, i.a.). We draw on these papers in our own investigation of how papers describe the goals of their work, the approaches they take in evaluating progress towards those goals, and the RAI issues they may raise.

2.2.2 Responsible AI in ATS

Task characteristics specific to ATS (e.g., reliance on source document as main source of information, output quality criteria specific to ATS) may foreground some RAI concerns more than others, and may even give rise to RAI concerns unique to ATS. As a result, it is just as important to study RAI in the context of specific tasks — in this thesis, the task of ATS. While a great deal of work on RAI issues has emerged for NLP broadly, much less work has addressed summarization specifically. Here, we provide an overview of such existing work.

Carenini and Cheung (2008) examine whether a summary reflects the distribution of opinions in customer reviews (i.e., multiple source documents) through an user study and a novel measure of opinion controversiality. Shandilya et al. (2018) and Dash et al. (2019) consider the notion of fairness in summarization, specifically focusing on Twitter data.Shandilya et al. (2018) studies whether extractive ATS systems fairly represent tweet authors from different genders (male v.s. female) and whether they fairly represent tweets expressing different political opinions (Democratic, Republican or neutral). Shandilya et al. (2020) explore readers' perceptions of fairness in summaries, finding that ROUGE metrics are not well-suited to capturing perceptions of summary fairness. Meanwhile, Keswani and Celis (2021) find that summarization systems produce summaries under-representing already-minoritized language varieties. In our analysis of the summarization literature, we explore to what extent papers acknowledge these existing concerns and aim to uncover issues not previously raised by existing work.

Chapter 3

Methodology

To understand research practices surrounding how, when, and which RAI issues are or should be considered by the ATS research community, we conducted a systematic survey of the ATS literature. To do so, we followed several steps which we elaborate in this chapter: we first i) gathered a collection of recent ATS papers to be examined and annotated (§3.1) and ii) reviewed a small set of ATS papers published in various venues to explore relevant practices (§3.2.1). Drawing on this exploratory review, iii) we then developed an annotation scheme (detailed in §3.3), which we used to annotate our collections of ATS papers (§3.2.2). Finally, iv) we analyze the annotations to understand emerging practices (§3.2.1).

3.1 Paper Collection

We focused on papers published between 2020 and 2022 in the ACL Anthology, a repository of papers maintained by the Association for Computational Linguistics (ACL). It currently hosts over 90K papers on the study of computational linguistics and NLP, including publications from major research conferences in NLP such as ACL (namesake conference of the Association) and Empirical Methods in Natural Language Processing. [1]

[1] https://aclanthology.org/

To collect papers on the task of ATS, we first gathered all papers with "summarization" in their title or abstract.[2] We then manually removed unrelated or irrelevant papers based on their abstracts. The set of unrelated papers consist of papers using the keyword of "summarization" for purposes other than the task of ATS (e.g., video summarization). The set of irrelevant papers consist of papers where the main focus was not ATS (e.g., natural language generation papers where ATS is one of many evaluated tasks). This resulted in the set of 333 papers that we annotate in §3.2.2.

3.2 Paper Review & Annotation

3.2.1 Exploratory Review

To scope our literature review and determine the practices we wanted to capture, we started with a small set of 8 ATS papers. We wanted a variety of papers in terms of publication venues and domain, and we were also interested in papers with an "ethical considerations" section. The selected papers (unrelated to the set collected in §3.1) are either published at ACL venues (DeYoung et al., 2021; Zhao et al., 2020; Feng et al., 2021; Aralikatte et al., 2021; Zhang et al., 2021b) or HCI and social computing venues (Zhang et al., 2020; Tran et al., 2020; Molenaar et al., 2020), and they cover summarization of medical literature, medical dialogues, legal cases, and emails. We observed differences among these papers, with those published at HCI/social computing venues focusing more on how ATS systems are used and on their stakeholders, which, along with our research questions, informed our early annotation aspects. These aspects included *mentioned stakeholders*, *author affiliation*, *domain*, *limitations*, and more.

3.2.2 Paper Annotation

From this starting point, we developed a common annotation scheme over several iterations (Rounds 1 & 2 below), which was then used to annotate the collection of ATS papers. Each paper took on average 23 minutes to annotate.

[2]We excluded demonstration, shared task description, tutorial, and workshop papers.

Round 1: Developing & refining the annotation scheme. Guided by the initial annotation dimensions mentioned in §3.2.1, every author open-coded 20 papers such that each paper was coded by 2 authors, totaling 60 papers. We periodically compared our annotations, updated the annotation scheme to resolve confusions and disagreements, and revised our annotations when necessary. For example, we split the initial *domain* category into *intended domain* and *actual domain* to better track differences between the two, which we noticed in many papers. At the end of this round, we arrived at the scheme overviewed in §3.3.

Round 2: Applying & clarifying the annotation scheme. Using the scheme from Round 1, we coded a larger subset of 131 papers. While in this round each paper was coded by a single author, we continued to periodically discuss ambiguous cases, clarify the annotation scheme, and update annotations accordingly.

Round 3: Hired annotators. With the guidelines finalized, for the remainder of the papers, we hired 7 annotators. These annotators are all graduate students in the field of NLP. We paid them at a rate of 30 CAD per hour, which is roughly equivalent to the wage of teaching assistants at our university. We started by briefing the annotators with a 1.5-hour paid training session on our project goal, how their annotations would be used, and the annotation scheme, illustrated by examples from the first two annotation rounds. We then scheduled 26 two-hour sessions held via video-conference, with one author present at all times to answer questions and offer clarifications. The annotators could choose which and how many sessions to attend, and were reminded of their right to periodically suspend or quit the annotation, without any impact on their pay. A total of 142 papers were annotated by hired annotators.

3.3 Annotation Scheme

To help us reflect on when RAI issues are brought up, how they are framed, and by whom, our annotation scheme covers aspects related to each paper's goals & authors (§3.3.1), evaluation practices (§3.3.2), as well as stakeholders (if any mentioned), limitations (of prior work or current work), and ethical considerations (§3.3.3).

3.3.1 Paper Authors & Goals

As we aim to examine how practitioners engage with RAI in their work, we need to know who the practitioners are, what their work is, and what motivates their work. These aspects not only contextualize our survey, but also provide cues about potential usage scenarios, which may determine what harms are likely to occur. Specifically, we consider the following aspects:

- **Contributions:** the type(s) of contribution a paper makes to the research community, including a new *dataset*, a ATS *system* (including new models, methods, or techniques), an evaluation *metric*, an *application* of ATS, a comprehensive *evaluation* of a collection of existing artifacts, or *other* types of contributions. This allows us to examine, for instance, whether authors of papers with certain types of contributions are more likely to engage in ethical reflection.

- **Intended domain:** the domain(s) the work is stated to be developed for, including *news* articles, *dialogue*, computer *code*, *medical* documents, *blogs* (e.g., Twitter), *opinions* (e.g., customer reviews), *scientific* articles, *wiki* (Wikipedia or Wikipedia-like platforms), *other* domains, or *general*. The latter code is used when nothing is explicitly specified throughout the paper's introduction (i.e., a failure to state an intended domain), or when the paper explicitly intends to be general (i.e., explicitly stating that its contribution is general-purpose, or that it can be used in any domain or application).

- **Research goals:** authors' stated goals. Annotators either copy or summarize the paper's goal, based on the abstract and the introduction of the paper. This provides additional context to the contributions, intended use or domain, as well as issues with current practices the authors aim to address.

- **Affiliation:** the authors' affiliation, including whether there is at least one author affiliated with an *academic* institution, with *industry*, or *other* organizations (e.g., government).

3.3.2 Data & Evaluation Practices

Evaluation practices reflect the space of concerns (including RAI issues) that the community is aware of, and can also give rise to their own RAI issues. We thus annotate papers according to:

- **Actual domain:** the domain(s) of the data that is actually used in the papers for evaluation or other purposes, with the same codes as the intended domain. This enables us to examine discrepancies between intended and actual domains.
- **Quality criteria:** text properties practitioners focus on when evaluating ATS systems. We annotate this aspect to understand what is conceptualized as a "good" summary. The annotators copy the name of properties (e.g., "factuality"), or passages describing the properties. When ROUGE or ROUGE-like automatic metrics are used without naming any particular target property, the annotators assign the keyword "ROUGE".

3.3.3 Limitations & Ethical Considerations

Lastly, we are also interested in both what kind of limitations (of both their work and prior work), ethical considerations, and stakeholders the authors explicitly bring up, as well as limitations that they might have overlooked.

- **Limitations of prior work:** what authors describe as weaknesses of prior work to track what existing issues the authors engage with. To capture this, annotators copy or summarize passages where limitations of prior work are covered.
- **Limitations of one's work:** whether and how the authors discuss the limitations of their own work. Annotators again either copy or summarize relevant passages.
- **Other limitations identified by annotators:** the notes annotators took about any limitations they noticed while reviewing that were not already mentioned by the authors.
- **Ethical considerations:** whether there is a paragraph or section dedicated to ethical considerations, broader impacts of the work, or similar related topics, and if so, what is discussed therein. Annotators copy or summarize the section.

- **Stakeholders:** whether the authors mention any stakeholders and who these stakeholders are, including *human annotators*, existing or anticipated *users* of a system, other *researchers* (in machine learning, AI, NLP, or related fields), or *other stakeholders*. We track this information because considering stakeholders is critical for envisioning harms and unintended consequences (Boyarskaya et al., 2020; Buçinca et al., 2023).

3.4 Analyzing Annotations

In addition to the codes we assigned to papers during the annotation, to further characterize particular practices (e.g., how many papers evaluate for factuality?), we also used keyword search and measured keyword frequency. To further assess various subsets of papers (e.g., papers that discuss their own limitations), we performed qualitative coding on extracted quotes, revisiting papers as necessary. Appendix 6.1 contains more details on this analysis (e.g., keywords used, codes for qualitative coding).

Chapter 4

Findings

Our systematic review surfaced insights related to what the ATS research community has been focusing on (§4.1), common evaluation practices the community employs (§4.2), and how the community engages with ethical considerations (§4.3).

4.1 Community Focus

To help unpack why and how authors might (or might not) approach RAI considerations, we first wanted to understand who is conducting the research and how they conceptualize their work—what they are creating, who they are creating it for, and what outcomes they envision—helping us to contextualize current and possible future practices.

There is a focus on developing new systems, with comparatively less emphasis on evaluation, metrics, datasets, or applications. Nearly 70% of the 333 papers contribute new *systems* (including models and methods), while fewer than 30% contribute new datasets and around 10% contributed new metrics. An even smaller fraction (less than 2%) of papers focus on applications. This emphasis on developing new ATS systems, models, or techniques echoes concerns about the devaluation of e.g., data work which is often framed as "peripheral, rather than central" to AI research Gero et al. (2023)—in contrast to the prestige of doing what is perceived to be more "technical" work such as modelling or system building.

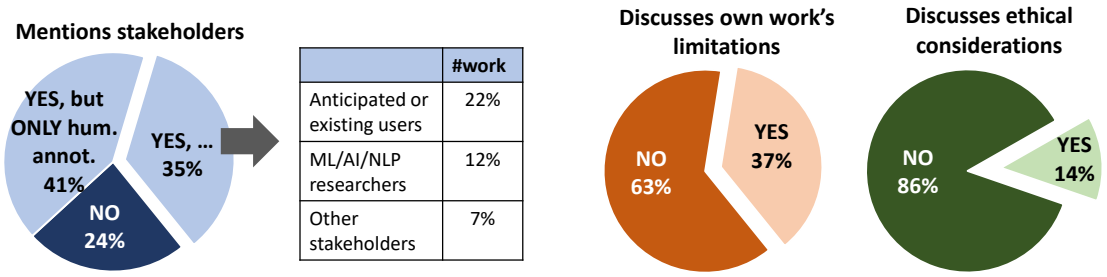

Figure 4.1: Summary statistics about mentioned stakeholders, and about how often papers cover limitations of one's work and ethical considerations.

Contributions	#	**Intended Domain**	#
System	224	General	173
Dataset	91	News	45
Metric(s)	36	Dialogue	38
Evaluation	73	Opinion	19
Application(s) & Other	34	Medical	17
Author Affiliation	#	Scientific	10
Academic	299	Code	9
Industry	121	Wiki	9
(collab of above two)	(95)	Blog	3
Other	32	Other	26

Table 4.1: Overview of the corpus of papers we reviewed in terms of contribution type, author affiliation and intended domain.

Many systems, metrics and datasets are intended to be general-purpose. Assuming that not explicitly stating an intended domain means the work is implicitly intended to be "general-purpose," ∼55% of 224 of papers contributing new systems intend them to be "general-purpose." Similarly, ∼72% (26 out of 36) of papers contributing metrics and ∼23% (21 out of 91) of those contributing datasets are also intended for general-purpose settings. However, these ostensibly general-purpose artifacts are often only tested or trained on a few domain-specific datasets and scenarios (§4.2).

The ATS literature remains driven by academic research, though there is also significant interest from industry. ∼90% of the reviewed papers were co-authored by at least one academia-affiliated author, with ∼32% of these papers being collaborations with industry authors. There are comparatively fewer papers solely written by authors affiliated with industry or other non-academic organizations. Because such organizations may be more connected to specific deployment settings and users Zhou et al. (2022), their low representation may represent a barrier to engaging with the impacts of ATS systems.

Papers rarely mention stakeholders when imagining intended use contexts. While ∼76% of reviewed papers mention stakeholders, fewer than half seem to mention stakeholders *other than* human annotators, who are typically mentioned in the context of evaluation practices rather than when discussing research goals. Only ∼22% of all annotated papers consider anticipated or existing users, while ∼12% mention other researchers. Conceptualizing a contribution without conceptualizing stakeholders may mean that the contribution will not meaningfully benefit any particular stakeholders, and may also make it more difficult to reason about limitations or adverse impacts.

Papers referencing users are often both explicit about who those users are and specify an intended domain. About three-fourths of the 74 papers we found to explicitly reference users, both describe who these users are and how they would benefit from ATS, e.g., *"automatic summarizing tool that can generate abstracts for scientific research papers [...] can save much time for researchers and also readers"* To et al. (2021). Many of these papers explicitly mention a specific intended domain, with only a small fraction of them (about 8%) (implicitly or explicitly) intending their work to be general-purpose. The remaining one-fourth only vaguely mention users, sometimes by specifying what users might want (e.g., *"a user might be looking for an overview summary or a more detailed one"* Xu and Lapata (2020)), without specifying who they might be. This is particularly the case when the authors intend for their work to be general purpose (67%, 12 out of 18). Not having a clear application or domain in mind can, however, make it difficult for authors to imagine users or other stakeholders.

Imagined benefits often only include reducing anticipated users' labor or improving customer experiences. ∼54% (40 out of 74) of papers referencing users aim to reduce some type of labor. In these instances, the work is meant to automate, speed up, or even replace parts of users' workflow. Examples include helping workers by summarizing meetings or emails (e.g., Singh et al., 2021; Zhang et al., 2022) and helping health professionals by summarizing medical encounters or files (e.g., Hu et al., 2022; Adams et al., 2021). ∼20% of 74 papers referencing users aim to improve customer experience, which involves a transactional relationship between those who would deploy the ATS system and those who would use output summaries, for example, summarizing product reviews to *"make the shopping process more useful and enjoyable for customers"* (Oved and Levy, 2021) or summarizing livestreamed content to *"fully meet the needs of customers [on livestreaming platforms]"* (Cho et al., 2021). A more expansive conception of benefits might help practitioners consider more stakeholders, applications, and impacts.

4.2 Evaluation Practices

To examine current practices we considered actual domains (i.e., the domain of the data used or collected in the paper), researchers' conceptualization of summary quality, and which quality criteria they tend to prioritize.

Most papers on general-purpose systems, metrics, and datasets *solely* use data from the *news* domain. We estimate that ∼52% of 122 papers contributing general-purpose systems only use *news* data when developing or evaluating systems, methods or models. This is not surprising since the most common ATS datasets are from the *news* domain (Dernoncourt et al., 2018; El-Kassas et al., 2021). For general-purpose metrics (i.e., not developed for only a restricted set of applications or domains and meant to be applied broadly), this percentage is ∼77% out of 26 papers. Similarly, datasets introduced by papers that do not state an intended domain, or explicitly aim to be general-purpose, all collect their data from the *news* domain. These practices could

introduce risks, as e.g., systems ostensibly developed to be general-purpose but only trained and evaluated on a restricted set of domains cannot be reliably used in other domains.

While quality criteria concerning information saliency, linguistic properties, and factuality are frequently evaluated, criteria such as bias and usefulness are rarely evaluated, if ever. Criteria related to information saliency (e.g., "informativeness," "relevance," "redundancy") are mentioned by ∼41% of all reviewed papers. This is followed by criteria related to linguistic properties (e.g., "coherence," "fluency," "readability"), mentioned by ∼39% of papers, and criteria related to factuality (e.g., "factual consistency," "hallucination," "faithfulness"), mentioned by ∼28% of papers. Other criteria, such as summary usefulness (e.g., *"how useful is the extracted summary to satisfy the given goal, in our case, to answer the given query"* (Iskender et al., 2020)), and whether the summaries exhibit some bias (e.g., bias in text sentiment polarity (Sarkhel et al., 2020)) are rarely if ever mentioned. As a consequence, current practice seldom assesses whether more user-facing goals of ATS (e.g., the actual reduction of labor) are attained. Task-based evaluation, where summaries are assessed based on how they help humans perform a particular task (Lloret et al., 2018), is not a foreign concept in ATS (Van Labeke et al., 2013; Zhu and Cimino, 2015; Jimeno-Yepes et al., 2013) and could be adopted by the community to better suit certain research goals.

While factuality, information saliency, and linguistic properties are frequently evaluated, these criteria are less commonly conceptualized as part of research goals and limitations. Comparatively, only ∼10% of all examined papers explicitly aim to address factuality-related qualities (e.g., better "evaluate faithfulness," "localizing factuality errors" in output summaries, or to prevent model "hallucinations"), and ∼15% of papers note factuality-related limitations in prior work (e.g., *"generating summaries that are faithful to the input is an unsolved problem"* (Aralikatte et al., 2021)). Similarly, only 8% of examined papers consider information saliency-related criteria as part of research goals, while 12% of papers point to these criteria when covering limitations of prior work. For linguistic properties, these percentages are 5% for research goals, respectively 9% for limitations of prior work. Naming these criteria as desirable, and explicitly targeting them in

research, would facilitate the adoption of more careful operationalizations and engagement with the risks they may give rise to.

Evaluation of output summaries still relies heavily on ROUGE or on other similar automatic metrics based on lexical overlap, with ∼90% of 224 papers proposing new systems using such metrics. This fraction is ∼87% (79 out of 91) for papers contributing datasets and ∼70% (51 out of 73) for papers providing comprehensive evaluations. Overall, about 22% of all examined papers *only* use these metrics. Since the reliability of ROUGE has been questioned (Novikova et al., 2017; Bhandari et al., 2020), there is a risk that metric scores do not reflect the true performance of evaluated systems.

Human evaluation is widely used, but details on how it is carried out are often missing. Some form of human evaluation seems used in a majority of papers, with ∼58% of all papers including mentions of human annotators. Yet our paper annotators noted limitations in how these evaluations were carried out for 24% (47 out of 194) of papers mentioning human annotators. Some of the issues most salient to our paper annotators included papers lacking detail about who the human evaluators are (noted for 22, ∼47% of 47 papers); the text properties or quality criteria human evaluators were asked to rate, such as asking annotators to score "importance" and "readability" without providing clear definitions (noted for 11, ∼23% of 47 papers); and the evaluation process in general, such as whether annotators were shown source documents during evaluation (noted for 9, ∼19% of 47 papers). These issues are particularly problematic for reproducibility and research standards. The community could adopt best practices developed for evaluation design, transparency, and analysis in human evaluation of text generation systems (van der Lee et al., 2019; Schoch et al., 2020).

4.3 Limitations and Ethical Considerations

Finally, we examine whether and how the community has engaged with ethical considerations and limitations of their own work and of existing work.

Most papers do not discuss the limitations of their own work, and rarely include any ethical reflections. We estimate that ∼63% of all annotated papers do not include a discussion about the limitations of their own work, while only ∼14% of surveyed papers have a section on ethical considerations. Papers proposing datasets are more likely to have an ethical considerations section (∼20%, 19 out of 92) than those proposing systems (∼10%, 23 out of 224). Work without such explicit reflections may not be able to effectively incorporate potential weaknesses or ethical concerns into the design and evaluation of their proposed systems, datasets, or metrics.

When authors conceptualize ethical concerns, they often turn to data-related issues. ∼62% of the 45 papers we found to include ethical considerations sections cover data issues in these sections. The data-related issues that are foregrounded include: data access and copyright (21 papers)—e.g., specifying that the data is publicly available; data privacy (13 papers)—mostly stakeholders who are either the people producing the data (e.g., professional writers of a scraped website (Liu et al., 2021)) or the people described by the data (e.g., users and customer service agents of e-commerce websites where data is collected (Lin et al., 2021)); and data "bias" (11 papers).

When mentioned, data bias remains poorly defined or under-specified. When discussing possible biases in their data, papers tend to only briefly and generically mention "bias" or a type of "bias" (e.g., "political bias", "gender bias", "biased views"). From our assessment, *only* 3 papers seem to provide more detail beyond these brief mentions (Adams et al., 2021; Cao and Wang, 2022; Zhong et al., 2021). Yet, even when bias issues are discussed in more depth, what is meant by data bias or the concerns or harms it can give rise to remain vaguely specified. For instance, (Zhong et al., 2021) mention how *"meeting datasets rarely contain any explicit gender information, [yet] annotators still tended to use 'he' as pronoun"* without further elaboration about e.g., the harmful stereotypes these biases might reproduce or whether the viewpoints of certain users might be unequally represented or misattributed in resulting systems' meeting notes summaries. While it is encouraging to see data bias identified as a source of concern, there is an opportunity to do so

consistently and to provide a clearly articulated conceptualization of what is meant by data bias (Blodgett et al., 2020a; Goldfarb-Tarrant et al., 2023).

While papers often discuss limitations related to various quality criteria, these are rarely conceptualized as ethical concerns. Papers describe a range of issues when reflecting on limitations. ∼24% of the 122 papers we found to discuss limitations talk about factuality-related issues—ranging from only brief mentions (e.g., generic references to "factual errors" or "hallucinations"), to more detailed descriptions (e.g., *"factual errors by mixing up important details [such as] mixing up the victim and suspect of a crime, mixing up locations and dates"* (Panthaplackel et al., 2022)). Limitations related to linguistic properties (e.g., length, word novelty, coherence, fluency) are also sometimes mentioned (20 out of 122), as are issues related to information saliency or coverage (12 out of 122).

Of these issues, however, only factuality seems to be conceptualized as an ethical concern, with ∼38% (17 out of 45) of papers with ethical consideration sections mentioning factuality-related concerns. From our assessment, no papers covering ethical concerns seem to relate them to quality criteria such as linguistic properties, or information saliency or coverage.

While factuality is sometimes conceptualized as an ethical issue, few papers reflect on the impact of factual errors. Only 6 of 17 papers (∼35%) seem to name factuality as an ethical concern by describing adverse impacts of factuality-related model failures, with 4 naming "misinformation" or "bad influence" in the news domain, one "misinformation" in the context of corporate meetings (e.g., which *"would negatively affect comprehension and further decision making"* (Zhong et al., 2021)), and one the *"risk of misinterpretation of evidence and subsequent [medical] malpractice"* (Otmakhova et al., 2022).

The other 11 papers either generically describe factuality-related model failures (e.g., *"Even though our models yield factually consistent summaries [...] they can still generate factually inconsistent summaries or sometimes hallucinate information"* (Jiang et al., 2021), or describe factuality-related concerns as an "open problem" (Xiao et al., 2022) or as a problem with "unacceptable outcome" in "high-impact" domains (such as scientific and medical domains, DeYoung

et al. (2021)) without much elaboration. A few papers (3) also explicitly warn that their models are not ready for deployment due to the lack of guarantees for the factual correctness of model outputs.

Papers describing ethical considerations often do not engage with intended use context. We estimate that fewer than half of the 45 papers explicitly considering ethical issues engage with intended use contexts. For example, only 3 papers explicitly mention the need for human oversight in system deployment, and only 2 of these describe the stakeholders who would be responsible for supervision with one paper noting that *"[t]he most natural application of this technology is not as a replacement for a human scribe, but as an assistant to one. By providing tools that aid a human scribe one can mitigate much of the risk of system failures, such as hallucination"* (Zhang et al., 2021a). Ethical considerations not grounded in use contexts may not be able to realistically anticipate adverse impacts.

When stakeholders are mentioned in ethical considerations, potential harm to them is often overlooked. Discussion of stakeholders is restricted to the compensation of human annotators (13 out of 45 papers), data privacy (13 out of 45 papers), and intended positive impacts on anticipated users (15 out of 45 papers). This may limit the conceptualization and evaluation of benefits and harms. The above requirement for human oversight, for instance, does not consider whether it might increase labor instead of reducing it, nor does it consider when or which stakeholders are well-equipped to supervise.

Chapter 5

Discussion

In this chapter, we highlight current research practices and offer recommendations for how the community can improve ATS research practices to be more responsible. We additionally present a (non-exhaustive) list of open-ended questions that practitioners could use as guidance.

5.1 Intended Use Context

Intended use contexts are often not well-described in the papers we surveyed. Many papers do not specify an intended domain, and few works mention stakeholders, such as existing or intended users, when imagining intended use contexts. Even when such stakeholders are mentioned, imagined benefits can be quite narrow in scope, as we find imagined benefits to intended users are mostly restricted to reducing labor or improving customer experience. As a recommendation, we encourage practitioners to ***conceptualize their contributions' intended use context by articulating, as much as possible, relevant stakeholders, intended domains, and potential benefits and adverse impacts to those stakeholders.*** More specifically, we invite practitioners to consider the following questions.

– **Q1: Who are the intended users of the work's contribution, and how would they benefit from it?** If the work involves building or improving an ATS system, a straightforward example of intended users would be the anticipated or current users of said system. Who exactly are

they, and what do they use the system for? Even when the work is intended to be "general-purpose", it may be helpful to imagine a few different user profiles and the interaction these users would have with the system. This question is also applicable to other types of contribution: for example, a work comprehensively evaluating a collection of existing artifacts could intend to alert fellow practitioners about weaknesses in these artifacts and point them towards future research directions.

– **Q2: Who else is impacted, directly or indirectly, by the intended use described in Q1? Who is involved in the research itself, or in the artifacts used in the work?** We recommend practitioners to familiarize themselves with various types of possible stakeholders beyond the intended users. Bender (2019)'s typology of the risks of adverse impacts of NLP technology, for instance, considers direct stakeholders *by choice* and *not by choice*, and three types of indirect stakeholders: *subjects of query*, *contributors to broad corpus*, and *subjects of stereotypes*. They are all relevant to ATS: for instance, a person deciding to use an ATS system to summarize news articles would be a direct stakeholder *by choice*, while an unaware visitor of a news website employing an ATS system for headline generation would be a direct stakeholder *not by choice*; people described in the summarized news articles (e.g., politicians, suspects and victims of crimes) could be considered *subjects of query*, while writers of these articles would be *contributors to broad corpus*; if an ATS system perpetuates stereotypes against certain groups through its output summaries, then these targeted groups would be *subjects of stereotypes*.
– **Q3: What kind of input is the system meant to receive? Conversely, is there any kind of input that the system is not designed to properly handle?** The answer could involve, for example, the number of documents (i.e., single vs. multi-document ATS), the length (e.g., long document ATS), the domain of application (e.g., summarizing papers from a specific field of science), and the authors of the documents (e.g., social media posts written by private vs. public figures). This question could be adapted to work that is not centered on an ATS system: for metrics, for instance, what kind of summaries is the metric meant to be used for?

5.2 Evaluation Practices

Current evaluation practices may not provide meaningful insights about systems' true performance. In terms of quality criteria of interest to the research community, we find that the priority in ATS evaluation is often on information saliency, linguistic properties, and factuality. While these quality criteria are important, other more stakeholder-centric criteria (e.g., social bias, usability) might be just as relevant and important to measure in order to anticipate systems' true performance and impacts. Yet, we find them to be largely overlooked by the research community.

Furthermore, our survey surfaced commonly adopted evaluation methods that could be problematic. Notably, we found a mismatch between intended and actual domain, where around half of "general-purpose" systems only use news data in training and testing. This mismatch could produce evaluation results that are not representative of systems' performance on texts from other domains than news, and that are thus not indicative of whether these systems are truly "general-purpose". We also found a heavy reliance on ROUGE-like metrics, despite many prior works exposing the shortcomings of such metrics for evaluating summarization.

Based on these findings, we encourage authors to ***consider more stakeholder-centric criteria, clearly define them*** (which may require grounding them in specific use contexts), and ***adopt evaluation practices that meaningfully capture them***. Practitioners are invited to ask:

– **Q4: Is it feasible (e.g., budget-wise) to measure the intended impact? If not, what measurable proxies are instead considered, and how do they approximate the intended impacts?** In other words, what is a "good" summary under the context previously described in Q1-2-3? For instance, it might be unfeasible to measure how labor-saving an ATS system is for doctors in the management of patient reports. Instead, it may be feasible to define a set of criteria that characterizes useful summaries in this context (e.g., considering doctors' information needs when defining "informativeness"). It may also be helpful to re-frame the question: what is a "bad" summary? This could help surface evaluation criteria related to the system's failure modes and its potential adverse impacts.

- **Q5: How are the criteria described in Q4 evaluated, and how do the employed evaluation methods meaningfully capture these criteria?**
  - Does the test data cover the range of expected input described in Q3?
  - If automatic metrics are used, is there some reasoning or empirical evidence supporting their appropriateness (i.e., that it is indeed appropriate to use them to evaluate the criteria)?
  - If human evaluation is performed, who are the human annotators, and how is the evaluation conducted (e.g., information given as context to evaluate a candidate summary)? Is there some reasoning or empirical evidence supporting the quality of human evaluation?

While these questions are targeted towards work that is centered on building or improving an ATS system, they are relevant to other types of contribution. For example, practitioners building a dataset for ATS could reflect on quality criteria implicitly present in the collected "gold" summaries. Finally, for practitioners to adopt better evaluation methods, these methods need to first exist. We thus encourage ***the development of evaluation instruments (e.g., benchmarks or human evaluation protocols), especially those tailored for (or adaptable to) specific use contexts***, so to provide ways for the community to gain more meaningful insights about systems' performance in their specific use contexts.

5.3 Limitations and Ethical Considerations

There is a lack of engagement with limitations and ethical concerns. We find that most papers do not have discussions on their own limitations, ethical considerations, and other related issues. When they do, they often focus on data-related concerns. This practice is not wrong by itself, but could be indicative of a narrow range of ethical concerns practitioners might be aware of. We especially highlight two areas that tend to be overlooked: i) some model failures are rarely conceptualized as ethical concerns; ii) intended use contexts, including stakeholders, are rarely involved in ethical considerations, which prevents authors from imagining potential harm to said stakeholders. To better engage with limitations and ethical concerns, we recommend ***reflecting***

explicitly on the decisions made throughout the research process. Going back to the previous five questions could help surface limitations of the work and potential adverse impacts:

- **Negative impacts through intended use and misuse**
  - (Q1:) Could intended users be negatively impacted by the work instead of reaping the intended benefits? If so, how? For example, could an ATS system meant to increase work efficiency instead increase users' cognitive load as they decide if they can trust its outputs?
  - (Q2:) Could the other stakeholders be negatively impacted? If so, how?
  - (Q1 & Q3:) What would happen if the system is used by people other than its intended users, or for purposes other than its intended use (e.g., malicious use)? What would happen if the system is used on documents that it cannot (or is not known to) properly handle?
- **Limitations of employed evaluation methods**
  - (Q4:) If it is infeasible to measure the intended impacts, what would happen if the measured criteria are not good enough proxies? Are there any other criteria that need to be measured, but that are presently impossible to measure? How do these two factors impact the interpretation of evaluation results and subsequent decision-making (e.g., concerning system deployment, usage)?
  - (Q4:) For each of the measured criteria, what would happen if output summaries performed poorly on it (e.g., the impact of an "uninformative" summary on the system users)? Additionally, imagine scenarios where output summaries "fail" on multiple criteria.
  - (Q5:) Do the employed evaluation methods lack reasoning or empirical evidence supporting their appropriateness? How does this impact the interpretation of evaluation results and subsequent decision-making (e.g., concerning system deployment, usage)?
  - (Q5:) Are there any evaluation results that are deemed unsatisfactory? Are these results further investigated and explained? Do they point to any potential weaknesses or system failure modes, and how could these impact stakeholders?

We encourage practitioners to imagine ways to address each surfaced limitation or potential adverse impact. For example, what information should be communicated to the intended users so that they are less likely to inadvertently misuse the ATS system? What are some future research di-

rections that the community could pursue in order to, for instance, obtain more empirical evidence about an employed metric or human evaluation protocol?

Furthermore, we encourage practitioners to ***engage with prior literature*** (e.g., Bender, 2019; Weidinger et al., 2022) on ethical concerns and real harms to which NLP systems can give rise, such as hate speech, stereotyping, and misinformation. This could help practitioners critically reflect on their own work and more clearly engage with issues that have already been recognized as ethical concerns in NLP, such as "bias" (Blodgett et al., 2020a). Practitioners could also familiarize themselves with ethical issues *the ATS community* has already recognized. Through our survey, we identified issues, such as the risk of misgendering stakeholders in ATS, which have already been pointed out by some members of the community e.g., (Zhong et al., 2021). Engaging with these issues could also help practitioners imagine limitations and ethical concerns of their own work.

Chapter 6

Conclusion

In this thesis, we studied how the ATS research community currently conceptualizes and engages with broader RAI issues, and discussed how this might be impacted by existing research practices. Through our survey of over 300 contemporary ATS papers from the ACL Anthology, we found that i) authors often do not specify the intended use context of their work, which could hinder their consideration of its societal impacts; ii) current evaluation practices, such as the choice of evaluated summary quality criteria, may not give practitioners meaningful insights about the systems' performance in real use settings; iii) few papers engage with limitations and ethical concerns, and when they do, such discussion often fail to involve key factors such as stakeholders. We thus conclude that there remains significant opportunity to foreground RAI in ATS research: we offer actionable guidance to encourage a reflective, collective research and reporting practice in ATS research and beyond.

6.1 Limitations and Future Work

Our findings are limited to the papers covered by our survey, which come from the ACL Anthology and are written in English. Works from other sources, such as venues with a different focus (e.g., venues focusing on AI applications) or those having a different demographic distribution than ACL, might paint a different picture. Future work could explore the differences and similarities

between different types of publication venues, in order to provide a more complete perspective on RAI in ATS. Future work could also expand beyond the task of ATS to cover other NLP tasks such as machine translation, question answering, conversation generation, etc. This future direction is especially relevant given the increasing popularity of language technologies and the growing space of tasks they are used for.

Our findings are also limited to the time period it covers: for example, between 2020-2022 some venues had not yet introduced the requirement to have a "Limitations" (or "Broader Impacts" or "Ethical Considerations") sections. Future work, it would be the picture might change in the future as such requirements become more and more standard. As future work, content analysis of papers throughout the years could provide insights on the effectiveness and impacts of the introduction of RAI related requirements.

Furthermore, our findings are limited by our paper annotation process. The annotation guidelines described in Section 3.3 could overlook attributes which we failed to imagine with our current understanding of RAI issues and the task of ATS. While we carried out the steps detailed in Section 3.2.2 with the goal of ensuring high annotation quality, the annotation process is imperfect. Although we have ensured that all annotators have NLP backgrounds and are all trained in our annotation scheme, not all of them have experience in RAI. They do not perfectly follow the annotation guidelines nor follow them the same way as a collective.

Additionally, our work only analyzes the content of the reviewed papers, thus focusing on how authors report their research. Studying other research outputs that accompany conference papers, such as released datasets and code repositories, would further enrich our understanding of RAI in ATS. For example, future work could examine whether or to what degree released datasets contain biases that the authors either acknowledged or overlooked in their papers, and how much privacy risks they truly pose to stakeholders (e.g., data annotators, people described in the data).

Finally, our recommendations are formulated based on our findings and are consequently also limited. Notably, the list of guiding questions is non-exhaustive and thus insufficient to help uncover all possible limitations and ethical issues within a work.

Ethical Considerations

Through this work, we intend to help NLP practitioners, especially those working on the task of ATS, better understand responsible AI issues under the context of this task. We also hope to encourage these practitioners to conduct more responsible ATS research by offering recommendations based on our findings. Our work could also be valuable to those wishing to better understand how various NLP communities (in this work, the ATS research community) engage with responsible AI. For example, it could help inform RAI researchers, policymakers, conference organizers, etc. in designing and implementing research guidelines related to responsible AI in NLP research.

Our work may have unintended negative impacts. For example, by foregrounding a limited set of ethical and other responsible AI concerns in our study, we may inadvertently suggest to readers that other issues deserve less consideration. This could mislead them into adopting and promoting research practices that disregard these issues — going against what we wish to accomplish with this work and potentially harming stakeholders of their work. It is thus important for readers to be aware of the scope and the limitations of our work detailed in Section 6.1. Notably, our findings and recommendations may not be applicable to other NLP tasks, nor generalizable to the field of NLP as a whole. Considering other related work (Section 2.2.1 and beyond) and conducting further studies (e.g., a similar survey for the task of machine translation) are necessary to paint a better and fuller picture of RAI in NLP.

Bibliography

The ICSI Meeting Corpus, vol. 1, http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1198793, 2003.

Adams, G., Alsentzer, E., Ketenci, M., Zucker, J., and Elhadad, N.: What's in a Summary? Laying the Groundwork for Advances in Hospital-Course Summarization, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4794–4811, Association for Computational Linguistics, Online, https://doi.org/10.18653/v1/2021.naacl-main.382, https://aclanthology.org/2021.naacl-main.382, 2021.

Angelidis, S. and Lapata, M.: Summarizing Opinions: Aspect Extraction Meets Sentiment Prediction and They Are Both Weakly Supervised, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, edited by Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., pp. 3675–3686, Association for Computational Linguistics, Brussels, Belgium, https://doi.org/10.18653/v1/D18-1403, https://aclanthology.org/D18-1403, 2018.

Aralikatte, R., Narayan, S., Maynez, J., Rothe, S., and McDonald, R.: Focus Attention: Promoting Faithfulness and Diversity in Summarization, 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), pp. 6078–6095, Association for Computational Linguistics, Online, https://doi.org/10.18653/v1/2021.acl-long.474, https://aclanthology.org/2021.acl-long.474, 2021.

Ashurst, C., Hine, E., Sedille, P., and Carlier, A.: Ai ethics statements: analysis and lessons learnt from neurips broader impact statements, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 2047–2056, 2022.

Ben Abacha, A., Shivade, C., and Demner-Fushman, D.: Overview of the MEDIQA 2019 Shared Task on Textual Inference, Question Entailment and Question Answering, in: Proceedings of the 18th BioNLP Workshop and Shared Task, edited by Demner-Fushman, D., Cohen, K. B., Ananiadou, S., and Tsujii, J., pp. 370–379, Association for Computational Linguistics, Florence, Italy, https://doi.org/10.18653/v1/W19-5039, https://aclanthology.org/W19-5039, 2019.

Bender, E. M.: A Typology of Ethical Risks in Language Technology with an Eye Towards Where Transparent Documentation Can Help, presented at The Future of Artificial Intelligence: Language, Ethics, Technology Workshop. https://bit.ly/2P9t9M6, 2019.

Benotti, L. and Blackburn, P.: Ethics consideration sections in natural language processing papers, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 4509–4516, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, https://doi.org/10.18653/v1/2022.emnlp-main.299, https://aclanthology.org/2022.emnlp-main.299, 2022.

Bhandari, M., Gour, P. N., Ashfaq, A., and Liu, P.: Metrics also Disagree in the Low Scoring Range: Revisiting Summarization Evaluation Metrics, in: Proceedings of the 28th International Conference on Computational Linguistics, pp. 5702–5711, International Committee on Computational Linguistics, Barcelona, Spain (Online), https://doi.org/10.18653/v1/2020.coling-main.501, https://aclanthology.org/2020.coling-main.501, 2020.

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H.: Language (Technology) is Power: A Critical Survey of "Bias" in NLP, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5454–5476, Association for Computational Linguistics, Online, https://doi.org/10.18653/v1/2020.acl-main.485, https://aclanthology.org/2020.acl-main.485, 2020a.

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H.: Language (Technology) is Power: A Critical Survey of "Bias" in NLP, in: ACL, 2020b.

Blodgett, S. L., Lopez, G., Olteanu, A., Sim, R., and Wallach, H.: Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets, 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), pp. 1004–1015, Association for Computational Linguistics, Online, https://doi.org/10.18653/v1/2021.acl-long.81, https://aclanthology.org/2021.acl-long.81, 2021.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, p. 4356–4364, Curran Associates Inc., Red Hook, NY, USA, 2016.

Boyarskaya, M., Olteanu, A., and Crawford, K.: Overcoming Failures of Imagination in AI Infused System Development and Deployment, in: Proceedings of the Navigating the Broader Impacts of AI Research Workshop, 2020.

Buçinca, Z., Pham, C. M., Jakesch, M., Ribeiro, M. T., Olteanu, A., and Amershi, S.: AHA!: Facilitating AI Impact Assessment by Generating Examples of Harms, arXiv preprint arXiv:2306.03280, 2023.

Caliskan, A., Bryson, J. J., and Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases, Science, 356, 183–186, https://doi.org/10.1126/science.aal4230, https://www.science.org/doi/abs/10.1126/science.aal4230, 2017.

Cao, M., Dong, Y., Wu, J., and Cheung, J. C. K.: Factual Error Correction for Abstractive Summarization Models, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 6251–6258, Association for Computational Linguis-

tics, Online, https://doi.org/10.18653/v1/2020.emnlp-main.506, https://aclanthology.org/2020.emnlp-main.506, 2020.

Cao, S. and Wang, L.: HIBRIDS: Attention with Hierarchical Biases for Structure-aware Long Document Summarization, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 786–807, Association for Computational Linguistics, Dublin, Ireland, https://doi.org/10.18653/v1/2022.acl-long.58, https://aclanthology.org/2022.acl-long.58, 2022.

Carenini, G. and Cheung, J. C. K.: Extractive vs. NLG-based Abstractive Summarization of Evaluative Text: The Effect of Corpus Controversiality, in: Proceedings of the Fifth International Natural Language Generation Conference, pp. 33–41, Association for Computational Linguistics, Salt Fork, Ohio, USA, https://aclanthology.org/W08-1106, 2008.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., and Raffel, C.: Extracting Training Data from Large Language Models, 2021.

Cho, S., Dernoncourt, F., Ganter, T., Bui, T., Lipka, N., Chang, W., Jin, H., Brandt, J., Foroosh, H., and Liu, F.: StreamHover: Livestream Transcript Summarization and Annotation, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 6457–6474, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, https://aclanthology.org/2021.emnlp-main.520, 2021.

Cohan, A. and Goharian, N.: Revisiting Summarization Evaluation for Scientific Articles, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), edited by Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., pp. 806–813, European Language Resources Association (ELRA), Portorož, Slovenia, https://aclanthology.org/L16-1130, 2016.

Dash, A., Shandilya, A., Biswas, A., Ghosh, K., Ghosh, S., and Chakraborty, A.: Summarizing User-Generated Textual Content: Motivation and Methods for Fairness in Algorithmic Summaries, Proc. ACM Hum.-Comput. Interact., 3, https://doi.org/10.1145/3359274, https://doi.org/10.1145/3359274, 2019.

Dernoncourt, F., Ghassemi, M., and Chang, W.: A Repository of Corpora for Summarization, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, https://aclanthology.org/L18-1509, 2018.

Devinney, H., Björklund, J., and Björklund, H.: Theories of" Gender" in NLP Bias ResearchTheories of Gender in Natural Language Processing, in: ACM FAccT Conference 2022, Conference on Fairness, Accountability, and Transparency, Hybrid via Seoul, Soth Korea, June 21-14, 2022, 2022.

DeYoung, J., Beltagy, I., van Zuylen, M., Kuehl, B., and Wang, L.: MSˆ2: Multi-Document Summarization of Medical Studies, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 7494–7513, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, https://aclanthology.org/2021.emnlp-main.594, 2021.

Dong, Y., Wang, S., Gan, Z., Cheng, Y., Cheung, J. C. K., and Liu, J.: Multi-Fact Correction in Abstractive Text Summarization, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 9320–9331, Association for Computational Linguistics, Online, https://doi.org/10.18653/v1/2020.emnlp-main.749, https://aclanthology.org/2020.emnlp-main.749, 2020.

El-Kassas, W. S., Salama, C. R., Rafea, A. A., and Mohamed, H. K.: Automatic text summarization: A comprehensive survey, Expert Systems with Applications, 165, 113 679, https://doi.org/https://doi.org/10.1016/j.eswa.2020.113679, https://www.sciencedirect.com/science/article/pii/S0957417420305030, 2021.

Fabbri, A., Rahman, F., Rizvi, I., Wang, B., Li, H., Mehdad, Y., and Radev, D.: 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), edited by Zong, C., Xia, F., Li, W., and Navigli, R., pp. 6866–6880, Association for Computational Linguistics, Online, https://doi.org/10.18653/v1/2021.acl-long.535, https://aclanthology.org/2021.acl-long.535, 2021a.

Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., and Radev, D.: Summeval: Re-evaluating summarization evaluation, Transactions of the Association for Computational Linguistics, 9, 391–409, 2021b.

Falke, T., Ribeiro, L. F. R., Utama, P. A., Dagan, I., and Gurevych, I.: Ranking Generated Summaries by Correctness: An Interesting but Challenging Application for Natural Language Inference, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2214–2220, Association for Computational Linguistics, Florence, Italy, https://doi.org/10.18653/v1/P19-1213, https://aclanthology.org/P19-1213, 2019.

Feng, X., Feng, X., Qin, L., Qin, B., and Liu, T.: Language Model as an Annotator: Exploring DialoGPT for Dialogue Summarization, 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), pp. 1479–1491, Association for Computational Linguistics, Online, https://doi.org/10.18653/v1/2021.acl-long.117, https://aclanthology.org/2021.acl-long.117, 2021.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., and Srikumar, M.: Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI, Berkman Klein Center Research Publication No. 2020-1, https://doi.org/10.2139/ssrn.3518482, 2020.

Ganesan, K., Zhai, C., and Han, J.: Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions, in: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 340–348, Association for Computational Linguistics, 2010.

Gero, K. I., Das, P., Dognin, P., Padhi, I., Sattigeri, P., and Varshney, K. R.: The incentive gap in data work in the era of large models, Nature Machine Intelligence, 2023.

Gkatzia, D. and Mahamood, S.: A Snapshot of NLG Evaluation Practices 2005 - 2014, in: Proceedings of the 15th European Workshop on Natural Language Generation (ENLG), pp. 57–60, Association for Computational Linguistics, Brighton, UK, https://doi.org/10.18653/v1/W15-4708, https://aclanthology.org/W15-4708, 2015.

Gliwa, B., Mochol, I., Biesek, M., and Wawer, A.: SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization, in: Proceedings of the 2nd Workshop on New Frontiers in Summarization, edited by Wang, L., Cheung, J. C. K., Carenini, G., and Liu, F., pp. 70–79, Association for Computational Linguistics, Hong Kong, China, https://doi.org/10.18653/v1/D19-5409, https://aclanthology.org/D19-5409, 2019.

Goldfarb-Tarrant, S., Ungless, E., Balkir, E., and Blodgett, S. L.: This prompt is measuring <mask>: evaluating bias evaluation in language models, in: Findings of the Association for Computational Linguistics: ACL 2023, pp. 2209–2225, Association for Computational Linguistics, Toronto, Canada, https://doi.org/10.18653/v1/2023.findings-acl.139, https://aclanthology.org/2023.findings-acl.139, 2023.

Graff, D., Kong, J., Chen, K., and Maeda, K.: English gigaword, Linguistic Data Consortium, Philadelphia, 4, 34, 2003.

Hahn, U. and Mani, I.: The challenges of automatic summarization, Computer, 33, 29–36, https://doi.org/10.1109/2.881692, 2000.

Hardmeier, C., Costa-jussà, M. R., Webster, K., Radford, W., and Blodgett, S. L.: How to Write a Bias Statement: Recommendations for Submissions to the Workshop on Gender Bias in NLP, arXiv preprint arXiv:2104.03026, 2021.

Hermann, K. M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P.: Teaching Machines to Read and Comprehend, in: NIPS, 2015.

Howcroft, D. M., Belz, A., Clinciu, M.-A., Gkatzia, D., Hasan, S. A., Mahamood, S., Mille, S., van Miltenburg, E., Santhanam, S., and Rieser, V.: Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions, in: Proceedings of the 13th International Conference on Natural Language Generation, pp. 169–182, Association for Computational Linguistics, Dublin, Ireland, https://aclanthology.org/2020.inlg-1.23, 2020.

Hu, J., Li, Z., Chen, Z., Li, Z., Wan, X., and Chang, T.-H.: Graph Enhanced Contrastive Learning for Radiology Findings Summarization, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 4677–4688, Association for Computational Linguistics, Dublin, Ireland, https://doi.org/10.18653/v1/2022.acl-long.320, https://aclanthology.org/2022.acl-long.320, 2022.

Iskender, N., Polzehl, T., and Möller, S.: Towards a Reliable and Robust Methodology for Crowd-Based Subjective Quality Assessment of Query-Based Extractive Text Summarization, in: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 245–253, European Language Resources Association, Marseille, France, https://aclanthology.org/2020.lrec-1.31, 2020.

Jiang, Y., Celikyilmaz, A., Smolensky, P., Soulos, P., Rao, S., Palangi, H., Fernandez, R., Smith, C., Bansal, M., and Gao, J.: Enriching Transformers with Structured Tensor-Product Representations for Abstractive Summarization, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4780–4793, Association for Computational Linguistics, Online, https://doi.org/10.18653/v1/2021.naacl-main.381, https://aclanthology.org/2021.naacl-main.381, 2021.

Jimeno-Yepes, A., Plaza, L., Mork, J. G., Aronson, A. R., and Díaz, A.: MeSH indexing based on automatically generated summaries, BMC Bioinformatics, 14, 208 – 208, 2013.

Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-Y., Mark, R. G., and Horng, S.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports, Scientific Data, 6, 317, 2019.

Keswani, V. and Celis, L. E.: Dialect Diversity in Text Summarization on Twitter, in: Proceedings of the Web Conference 2021, WWW '21, p. 3802–3814, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3442381.3450108, https://doi.org/10.1145/3442381.3450108, 2021.

Krishna, K., Bigham, J., and Lipton, Z. C.: Does Pretraining for Summarization Require Knowledge Transfer?, in: Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 3178–3189, Association for Computational Linguistics, Punta Cana, Dominican Republic, https://aclanthology.org/2021.findings-emnlp.273, 2021.

Kryscinski, W., McCann, B., Xiong, C., and Socher, R.: Evaluating the Factual Consistency of Abstractive Text Summarization, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 9332–9346, Association for Computational Linguistics, Online, https://doi.org/10.18653/v1/2020.emnlp-main.750, https://aclanthology.org/2020.emnlp-main.750, 2020.

Kumar, K. and Cheung, J. C. K.: Understanding the Behaviour of Neural Abstractive Summarizers using Contrastive Examples, 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), pp. 3949–3954, Association for Computational Linguistics, Minneapolis, Minnesota, https://doi.org/10.18653/v1/N19-1396, https://aclanthology.org/N19-1396, 2019.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L.: BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, edited by Jurafsky, D., Chai, J.,

Schluter, N., and Tetreault, J., pp. 7871–7880, Association for Computational Linguistics, Online, https://doi.org/10.18653/v1/2020.acl-main.703, https://aclanthology.org/2020.acl-main.703, 2020.

Lin, C.-Y.: ROUGE: A package for automatic evaluation of summaries, in: Text summarization branches out, pp. 74–81, 2004.

Lin, H., Ma, L., Zhu, J., Xiang, L., Zhou, Y., Zhang, J., and Zong, C.: CSDS: A Fine-Grained Chinese Dataset for Customer Service Dialogue Summarization, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 4436–4451, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, https://aclanthology.org/2021.emnlp-main.365, 2021.

Liu, F. and Liu, Y.: Correlation between ROUGE and Human Evaluation of Extractive Meeting Summaries, in: Proceedings of ACL-08: HLT, Short Papers, edited by Moore, J. D., Teufel, S., Allan, J., and Furui, S., pp. 201–204, Association for Computational Linguistics, Columbus, Ohio, https://aclanthology.org/P08-2051, 2008.

Liu, S., Chen, S., Uyttendaele, X., and Roth, D.: MultiOpEd: A Corpus of Multi-Perspective News Editorials, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4345–4361, Association for Computational Linguistics, Online, https://doi.org/10.18653/v1/2021.naacl-main.344, https://aclanthology.org/2021.naacl-main.344, 2021.

Lloret, E., Plaza, L., and Aker, A.: The challenging task of summary evaluation: an overview, Language Resources and Evaluation, 52, 101–148, 2018.

Ludec, C. L., Cornet, M., and Casilli, A. A.: The problem with annotation. Human labour and outsourcing between France and Madagascar, Big Data & Society, 10, 20539517231188 723, https://doi.org/10.1177/20539517231188723, https://doi.org/10.1177/20539517231188723, 2023.

McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., Post, W., Reidsma, D., and Wellner, P.: The AMI meeting corpus, in: Proceedings of Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research, edited by Noldus, L., Grieco, F., Loijens, L., and Zimmerman, P., pp. 137–140, Noldus Information Technology, 2005.

Michalopoulos, G., Williams, K., Singh, G., and Lin, T.: MedicalSum: A Guided Clinical Abstractive Summarization Model for Generating Medical Reports from Patient-Doctor Conversations, in: Findings of the Association for Computational Linguistics: EMNLP 2022, edited by Goldberg, Y., Kozareva, Z., and Zhang, Y., pp. 4741–4749, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, https://doi.org/10.18653/v1/2022.findings-emnlp.349, https://aclanthology.org/2022.findings-emnlp.349, 2022.

Molenaar, S., Maas, L., Burriel, V., Dalpiaz, F., and Brinkkemper, S.: Medical Dialogue Summarization for Automated Reporting in Healthcare, pp. 76–88, https://doi.org/10.1007/978-3-030-49165-9_7, 2020.

Mullenbach, J., Pruksachatkun, Y., Adler, S., Seale, J., Swartz, J., McKelvey, G., Dai, H., Yang, Y., and Sontag, D.: CLIP: A Dataset for Extracting Action Items for Physicians from Hospital Discharge Notes, 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), pp. 1365–1378, Association for Computational Linguistics, Online, https://doi.org/10.18653/v1/2021.acl-long.109, https://aclanthology.org/2021.acl-long.109, 2021.

Nanayakkara, P., Hullman, J., and Diakopoulos, N.: Unpacking the expressed consequences of AI research in broader impact statements, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 795–806, 2021.

Narayan, S., Cohen, S. B., and Lapata, M.: Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization, ArXiv, abs/1808.08745, 2018.

Nedoluzhko, A., Singh, M., Hledíková, M., Ghosal, T., and Bojar, O.: ELITR Minuting Corpus: A Novel Dataset for Automatic Minuting from Multi-Party Meetings in English and Czech, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, edited by Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., pp. 3174–3182, European Language Resources Association, Marseille, France, https://aclanthology.org/2022.lrec-1.340, 2022.

Novikova, J., Dušek, O., Cercas Curry, A., and Rieser, V.: Why We Need New Evaluation Metrics for NLG, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2241–2252, Association for Computational Linguistics, Copenhagen, Denmark, https://doi.org/10.18653/v1/D17-1238, https://aclanthology.org/D17-1238, 2017.

Otmakhova, Y., Verspoor, K., Baldwin, T., and Lau, J. H.: The patient is more dead than alive: exploring the current state of the multi-document summarisation of the biomedical literature, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 5098–5111, Association for Computational Linguistics, Dublin, Ireland, https://doi.org/10.18653/v1/2022.acl-long.350, https://aclanthology.org/2022.acl-long.350, 2022.

Oved, N. and Levy, R.: PASS: Perturb-and-Select Summarizer for Product Reviews, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 351–365, Association for Computational Linguistics, Online, https://doi.org/10.18653/v1/2021.acl-long.30, https://aclanthology.org/2021.acl-long.30, 2021.

Panthaplackel, S., Benton, A., and Dredze, M.: Updated Headline Generation: Creating Updated Summaries for Evolving News Stories, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 6438–6461, Association for Computational Linguistics, Dublin, Ireland, https://doi.org/10.18653/v1/2022.acl-long.446, https://aclanthology.org/2022.acl-long.446, 2022.

Rowe, N.: 'It's destroyed me completely': Kenyan moderators decry toll of training of AI models, The Guardian, https://www.theguardian.com/technology/2023/aug/02/ai-chatbot-training-human-toll-content-moderator-meta-openai.

Sarkhel, R., Keymanesh, M., Nandi, A., and Parthasarathy, S.: Interpretable Multi-headed Attention for Abstractive Summarization at Controllable Lengths, in: Proceedings of the 28th International Conference on Computational Linguistics, pp. 6871–6882, International Committee on Computational Linguistics, Barcelona, Spain (Online), https://doi.org/10.18653/v1/2020.coling-main.606, https://aclanthology.org/2020.coling-main.606, 2020.

Schoch, S., Yang, D., and Ji, Y.: "This is a Problem, Don't You Agree?" Framing and Bias in Human Evaluation for Natural Language Generation, in: Proceedings of the 1st Workshop on Evaluating NLG Evaluation, pp. 10–16, Association for Computational Linguistics, Online (Dublin, Ireland), https://aclanthology.org/2020.evalnlgeval-1.2, 2020.

Shandilya, A., Ghosh, K., and Ghosh, S.: Fairness of Extractive Text Summarization, in: Companion Proceedings of the The Web Conference 2018, WWW '18, p. 97–98, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, https://doi.org/10.1145/3184558.3186947, https://doi.org/10.1145/3184558.3186947, 2018.

Shandilya, A., Dash, A., Chakraborty, A., Ghosh, K., and Ghosh, S.: Fairness for Whom? Understanding the Reader's Perception of Fairness in Text Summarization, in: 2020 IEEE International Conference on Big Data (Big Data), pp. 3692–3701, https://doi.org/10.1109/BigData50022.2020.9378095, 2020.

Singh, M., Ghosal, T., and Bojar, O.: An Empirical Performance Analysis of State-of-the-Art Summarization Models for Automatic Minuting, in: Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation, pp. 50–60, Association for Computational Linguistics, Shanghai, China, https://aclanthology.org/2021.paclic-1.6, 2021.

Strubell, E., Ganesh, A., and McCallum, A.: Energy and Policy Considerations for Deep Learning in NLP, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, edited by Korhonen, A., Traum, D., and Màrquez, L., pp. 3645–3650, Association for Computational Linguistics, Florence, Italy, https://doi.org/10.18653/v1/P19-1355, https://aclanthology.org/P19-1355, 2019.

Tan, R. and Cabato, R.: Behind the AI boom, an army of overseas workers in 'digital sweatshops', The Washington Post, https://www.washingtonpost.com/world/2023/08/28/scale-ai-remotasks-philippines-artificial-intelligence/.

To, H. Q., Nguyen, K. V., Nguyen, N. L.-T., and Nguyen, A. G.-T.: Monolingual vs multilingual BERTology for Vietnamese extractive multi-document summarization, in: Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation, pp. 692–699, Association for Computational Linguistics, Shanghai, China, https://aclanthology.org/2021.paclic-1.73, 2021.

Tran, V., Le Nguyen, M., Tojo, S., and Satoh, K.: Encoded summarization: summarizing documents into continuous vector space for legal case retrieval, Artificial Intelligence and Law, 28, 441–467, 2020.

van der Lee, C., Gatt, A., van Miltenburg, E., Wubben, S., and Krahmer, E.: Best practices for the human evaluation of automatically generated text, in: Proceedings of the 12th International Conference on Natural Language Generation, pp. 355–368, Association for Computational Linguistics, Tokyo, Japan, https://doi.org/10.18653/v1/W19-8643, https://aclanthology.org/W19-8643, 2019.

Van Labeke, N., Whitelock, D., Field, D., Pulman, S., and Richardson, J.: OpenEssayist: Extractive summarisation and formative assessment of free-text essays., 2013.

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., Haas, J., Legassick, S., Irving, G., and Gabriel, I.: Taxonomy of Risks Posed by Language Models, FAccT '22, p. 214–229, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3531146.3533088, https://doi.org/10.1145/3531146.3533088, 2022.

Xiao, W., Beltagy, I., Carenini, G., and Cohan, A.: PRIMERA: Pyramid-based Masked Sentence Pre-training for Multi-document Summarization, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 5245–5263, Association for Computational Linguistics, Dublin, Ireland, https://doi.org/10.18653/v1/2022.acl-long.360, https://aclanthology.org/2022.acl-long.360, 2022.

Xu, Y. and Lapata, M.: Coarse-to-Fine Query Focused Multi-Document Summarization, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 3632–3645, Association for Computational Linguistics, Online, https://doi.org/10.18653/v1/2020.emnlp-main.296, https://aclanthology.org/2020.emnlp-main.296, 2020.

Zhang, K., Chen, J., and Yang, D.: Focus on the Action: Learning to Highlight and Summarize Jointly for Email To-Do Items Summarization, in: Findings of the Association for Computational Linguistics: ACL 2022, pp. 4095–4106, Association for Computational Linguistics, Dublin, Ireland, https://doi.org/10.18653/v1/2022.findings-acl.323, https://aclanthology.org/2022.findings-acl.323, 2022.

Zhang, L., Negrinho, R., Ghosh, A., Jagannathan, V., Hassanzadeh, H. R., Schaaf, T., and Gormley, M. R.: Leveraging Pretrained Models for Automatic Summarization of Doctor-Patient Conversations, in: Findings of the Association for Computational Linguistics: EMNLP 2021,

pp. 3693–3712, Association for Computational Linguistics, Punta Cana, Dominican Republic, https://aclanthology.org/2021.findings-emnlp.313, 2021a.

Zhang, S., Celikyilmaz, A., Gao, J., and Bansal, M.: EmailSum: Abstractive Email Thread Summarization, 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), pp. 6895–6909, Association for Computational Linguistics, Online, https://doi.org/10.18653/v1/2021.acl-long.537, https://aclanthology.org/2021.acl-long.537, 2021b.

Zhang, X., Geng, P., Zhang, T., Lu, Q., Gao, P., and Mei, J.: Aceso: PICO-Guided Evidence Summarization on Medical Literature, IEEE Journal of Biomedical and Health Informatics, 24, 2663–2670, https://doi.org/10.1109/JBHI.2020.2984704, 2020.

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., and Chang, K.-W.: Gender Bias in Contextualized Word Embeddings, 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), edited by Burstein, J., Doran, C., and Solorio, T., pp. 629–634, Association for Computational Linguistics, Minneapolis, Minnesota, https://doi.org/10.18653/v1/N19-1064, https://aclanthology.org/N19-1064, 2019.

Zhao, Z., Cohen, S. B., and Webber, B.: Reducing Quantity Hallucinations in Abstractive Summarization, in: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 2237–2249, Association for Computational Linguistics, Online, https://doi.org/10.18653/v1/2020.findings-emnlp.203, https://aclanthology.org/2020.findings-emnlp.203, 2020.

Zhong, M., Yin, D., Yu, T., Zaidi, A., Mutuma, M., Jha, R., Awadallah, A. H., Celikyilmaz, A., Liu, Y., Qiu, X., and Radev, D.: 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, pp. 5905–5921,

Association for Computational Linguistics, Online, https://doi.org/10.18653/v1/2021.naacl-main.472, https://aclanthology.org/2021.naacl-main.472, 2021.

Zhou, K., Blodgett, S. L., Trischler, A., Daumé III, H., Suleman, K., and Olteanu, A.: Deconstructing NLG Evaluation: Evaluation Practices, Assumptions, and Their Implications, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 314–324, Association for Computational Linguistics, Seattle, United States, https://doi.org/10.18653/v1/2022.naacl-main.24, https://aclanthology.org/2022.naacl-main.24, 2022.

Zhu, X. and Cimino, J. J.: Clinicians' evaluation of computer-assisted medication summarization of electronic medical records, Comput. Biol. Med., 59, 221–231, 2015.

Appendix A: Statistics on Paper Annotators

Annotators took on average around 23 minutes per paper. Excluding the exploratory review and Round 1, the number of papers coded by each annotator is:

Round 2 totalling 131 papers

- Author-Annotator 1: 15
- Author-Annotator 2: 20
- Author-Annotator 3: 5
- Author-Annotator 4: 2
- Author-Annotator 5: 84
- Author-Annotator 6: 5

Round 3 totalling 142 papers

- Hired Annotator 1: 12
- Hired Annotator 2: 14
- Hired Annotator 3: 21
- Hired Annotator 4: 42
- Hired Annotator 5: 42
- Hired Annotator 6: 2
- Hired Annotator 7: 9

While inter-annotator agreement (IAA) is often used as a proxy for annotation quality, particularly when trying to determine some ground truth, our paper reviews are not meant as a gold standard, and instead we looked to build consensus on ambiguous cases. We aim to surface issues in the current practices and provide rough estimates of how prevalent these issues might be. To ensure quality, we recruited annotators with expertise in the field, and encouraged frequent discussions of ambiguous cases.

Code	Theme description	Num. Papers
#annotator_pay	considerations related to how the annotators were paid	13
#annotator_description	demographic and other information related to the expertise of annotators	5
#data_access	how the data proposed/used can be accessed	21
#data_bias	concerns about biases in datasets	12
#stakeholder_privacy	privacy concerns related to stakeholders' information/data	13
#compute_info	concerns related to computational time and resources	4
#factuality	concerns related to factual errors, e.g.,"hallucinations"	17
#intended_use	describe the intended use domain, or the broader context	21
#no_deployment	explicitly mentions that deployment is too premature, not ready	4

Table A1: Resulting codes and corresponding themes in "ethical considerations" sections.

Code	Theme description	Num.
#novel_words	issues related to systems inability to use words outside of the input text vocabulary	2
#length	considerations related to the length of the input text or of the summary	9
#misgender	the output summaries referencing the wrong gender pronouns/terms	3
#relevance	whether the information contained in the output is relevant, non-redundant	8
#info_recall	whether key information from the source is included by the output	7
#factuality	aspects related to factual consistency, e.g.,"hallucinations"	29
#readability	properties related to e.g., coherence, fluency, grammar	12
#failure_mode	specific circumstances where model/metric/etc. perform poorly	12
#doubt_generalize	concerns about generalizability to other domains, languages, etc.	13
#weak_methods	known or potential weaknesses with their methodology	29
#weak_experiment	known or potential weaknesses with their experimental design	26
#complex_use	using a system or method is complex or requires extensive computational resources	7

Table A2: Resulting codes and corresponding themes in authors' discussions of the limitations of their own work.

Appendix B: Methodology

Here, we provide additional details about the protocols we followed while coding and analyzing the set of papers included in our review.

B1 Community Focus

To examine *research goals*, our analysis considered *mentioned stakeholders* as we were interested in how the authors envision anticipated or existing users to benefit from their work, and how these users are described. For this, we first identified the papers that mention users. As we observed that the code *other stakeholders* was sometimes used to denote users, we also manually filtered all

papers coded *other stakeholders* for mentions of potential or existing users. When the description of research goals in these papers did not mention users, we revisited the papers to locate passages elaborating on how users benefit, which we then iteratively coded to identify the themes covered in Section 4.1.

To check whether commonly evaluated quality criteria such as factuality, information saliency, and linguistic properties were also conceptualized as part of research goals, we used the same keywords listed in the next section (§B2) to estimate the number of papers focusing on these criteria (discussed in §4.2).

B2 Evaluation Practices

To surface insights about current evaluation practices, we primarily examined aspects related to the *actual domain*, as well as commonly considered *quality criteria*. For *actual domain*, we were interested in discrepancies with what the *intended domain* was meant to be. For *quality criteria*, we examined the words authors frequently use to describe the quality criteria they consider, and performed the following keyword searches to estimate how often authors consider these criteria:

– information coverage: "relevan" (for relevant/relevance), "repetition", "informat" (for information/informativeness), "redundancy", "salien" (for salient/saliency), and "content coverage"
– information presentation: "fluen" (for fluent, fluency), "gramma" (for grammar, grammaticality), "readab" (for readable, readability), "coheren" (for coherent, coherence), "length", "novel" (for novel words)
– factuality: "factual" (also for factuality), hallucinat (for hallucinate, hallucination), faithful (also for faithfulness), consisten (for consistency), correct (also for correctness).

To estimate how frequently ROUGE-like automatic metrics are used, was tracked it using the tag "ROUGE" during the paper annotations.

B3 Ethical Considerations

After inspecting the annotators' summaries provided by the annotators for the ethical consideration sections, we discarded 2 papers which we found to be mistakenly annotated as having an ethical considerations section: i) Krishna et al. (2021): the annotation pulled passages from the abstract. There's no ethical considerations section in the paper. ii) Mullenbach et al. (2021): the paper has a "potential impact" section in the introduction that we believe addresses the paper goal.

The specific codes we obtained after iteratively coding the ethical considerations sections to surface themes are listed in Table A1.

B4 Limitations of one's own work

The specific codes we obtained after iteratively coding the passages about the authors discussions of the limitations of their own work are listed in Table A2.