Beyond Paraphrasing: Analyzing Summarization Abstractiveness and Reasoning

Nathan Zeweniuk[1,2], **Ori Ernst**[1,2], **and Jackie Chi Kit Cheung**[1,2,3]

[1]Mila – Quebec Artificial Intelligence Institute
[2]McGill University
[3]Canada CIFAR AI Chair, Mila
{oriern}@gmail.com
{nathan.zeweniuk@mail., jackie.cheung@}mcgill.ca

Abstract

While there have been many studies analyzing the ability of LLMs to solve problems through reasoning, their application of reasoning in summarization remains largely unexamined. This study explores whether reasoning is essential to summarization by investigating three questions: (1) Do humans frequently use reasoning to generate new summary content? (2) Do summarization models exhibit the same reasoning patterns as humans? (3) Should summarization models integrate more complex reasoning abilities? Our findings reveal that while human summaries often contain reasoning-based information, system-generated summaries rarely contain this same information. This suggests that models struggle to effectively apply reasoning, even when it could improve summary quality. We advocate for the development of models that incorporate deeper reasoning and abstractiveness, and we release our annotated data to support future research.

1 Introduction

In recent decades, the amount of textual information available has grown exponentially, creating a pressing need for automatic systems that can process this information and derive meaningful conclusions from it. Recent advances in large language models (LLMs) have shown remarkable progress in handling tasks that appear to require reasoning-namely, deriving conclusions not explicitly stated in the text. For instance, LLMs have demonstrated strong performance in question answering tasks that involve background knowledge and inference (Zhao et al., 2023; Liu et al., 2025). Yet, despite these advances, the role of reasoning in generic summarization remains largely underexplored. A key question arises: Can and should automatic summaries incorporate new conclusions that go beyond the information explicitly present in the source?

Traditionally, research in automatic summarization has focused on information selection and paraphrasing (Zhang et al., 2018; Lebanoff et al., 2019b; Ernst et al., 2022). One widely used quality measure to human-like summaries has been abstractiveness-the degree to which a summary uses its "own words" rather than copying source text. With the emergence of LLMs, summarization systems have achieved substantial gains not only in content selection but also in producing highly fluent and abstractive outputs comparable to human-written summaries (Goyal et al., 2022). These advances invite a deeper investigation into the next frontier: Can automatic summaries perform reasoning, deriving conclusions like humans do?

In principle, the ability to reason during summarization could enhance content focus and informativeness, enabling the generation of summaries that emphasize the most salient insights rather than merely restating information. Table 1 illustrates how different summarization processes can yield outputs with varying levels of focus, conciseness, and reasoning.

To investigate this, we outline several research questions: (1) Do *humans* rely on reasoning to create summary content, and if so, how often? (2) Do *models* employ the same reasoning as humans, or do they bypass it? (3) Should we aim to incorporate reasoning abilities in summarization modeling?

To address these questions, we began with a manual annotation of human-generated summaries. We identified three common operations that humans perform when rewriting selected text for summaries (will be defined clearly in Section 3): paraphrasing, generalization, and drawing conclusions. The latter two operations, generalization and conclusion, change the semantic meaning from the source to the summary and are considered to require reasoning.[1] We manually classified matching text spans between summaries and source texts according to

[1]We acknowledge that reasoning may be used in paraphrasing, but such reasoning does not fall within the scope of our work as it does not lead to new semantic content.

Proceedings of The 5th New Frontiers in Summarization Workshop, pages 48–58
November 8, 2025 ©2025 Association for Computational Linguistics

Type	Text
Source	Investigators are trying to piece together what led to the deaths
Paraphrase	Investigators are attempting to figure out how the deaths occurred
Generalization	Investigators are working on a case
Conclusion	Mystery surrounds the death of two brothers

Table 1: An example of abstractiveness levels that can be applies to a source sentence

these levels of abstractiveness and found that approximately 25% of human summary spans involve such reasoning.

However, our manual evaluation of system-generated summaries revealed a different pattern. Despite high overall evaluation scores with respect to reference summaries, these systems predominantly matched the reference with paraphrased text spans. Crucially, reference spans that require reasoning to extract important information were underrepresented in systems summaries. In other words, while these models seem to display reasoning abilities in other tasks, they tend to avoid using them correctly in summarization, where output is often deemed acceptable without it, despite the fact that this omission can reduce the quality of the summary.

This analysis highlights the importance of reasoning in summarization and calls for the development of new models that better integrate reasoning and different levels of abstraction. To facilitate further research, we are releasing the manual annotations and data.[2]

2 Related Work

2.1 Abstractiveness Analysis

Prior research on abstractiveness has focused on how multiple source sentences are fused into one summary sentence (Barzilay and McKeown, 2005). Early work analyzed syntactic aspects of fusion, investigating whether sentences are merged or concatenated (Lebanoff et al., 2019a), and explored points of correspondence between sentences (Lebanoff et al., 2020). These studies were conducted on nearly-extractive data and based mostly on paraphrasing.

More recent research has expanded to include specific cases of entity generalization in summaries that went beyond paraphrasing (González et al., 2022; Jumel et al., 2020). A new aggregation metric (He et al., 2023) has also been introduced to capture even non-paraphrasing forms of fusion. Overall, most existing studies focus on sentence fusion and lack detailed manual annotations of abstractiveness across varying levels of sentence complexity.

The work most closely related to our approach is that of Jing (2002), who conducted a seminal study examining the operations employed by human summarizers when composing abstractive summaries. Specifically, they identified six key actions: sentence reduction, sentence combination, syntactic transformation, lexical paraphrasing, generalization/specification, and content reordering. Based on the hypothesis that abstractive summarization relies on these operations, one can characterize a summary's level of abstractiveness by measuring the extent to which each action is applied.

As mentioned, while modern summarization systems are capable of performing many of these operations, especially those related to paraphrasing, our work focuses on a different dimension: reasoning. We argue that, despite advancements in paraphrasing and surface-level transformations, reasoning-based abstraction, the ability to derive new conclusions or implicit insights, remains underexplored and is largely absent from prior work.

2.2 LLMs Reasoning Ability

Reasoning benchmarks, such as commonsense (Talmor et al., 2019), logical (Sinha et al., 2019), math (Saxton et al., 2019), or multi-hop (Tu et al., 2019), were considered a difficult task for language models to solve. Most of these benchmarks were designed in a Question-Answering (QA) setup, where a query or a question is given, and the answer can be found in a defined set of sources. In order to find the answer, the system is expected to use reasoning. Recent advancements in LLMs showed significantly improvement on these benchmarks (Driess et al., 2023; Touvron et al., 2023; Espejel et al., 2023), especially while using chain of thought (Kojima et al., 2022; Wei et al., 2022). However, all these benchmarks are designed to elicit reasoning. In contrast, this paper aims to evaluate the reasoning abilities of models in summarization tasks, where acceptable outputs can still be achieved with-

[2]Annotated data is publicly available at https://github.com/oriern/ReasoningSummarization.

out reasoning.

3 Abstractiveness Levels

There are a few ways in which source information can be utilized in the process of summarization. The simplest approach is to copy the information directly and reword or restructure it to fit the rest of the summary. Generalizing certain parts of the source material can help compress information even further by reducing the amount of detail and level of specificity. Sometimes, it is necessary to add entirely new information drawn from the source through reasonable conclusions to avoid relying on reader inference. We use these observations regarding the various uses of source information to define abstractiveness levels.

First, we define a span-level matching between the information in the summary and its corresponding evidence from the source. Having these matching pairs allow to analyze the abstractiveness level performed in the summary. Following Ernst et al. (2021, 2024) the spans are standalone facts that are usually formed into a proposition, where the source span entails the summary span. We also required *tight* matching, where each source token that adds additional information that the summary is not based on, is omitted.

Given these pairs, the abstractiveness levels are defined as follows:

Paraphrase. Bi-directional entailment between the summary span and the document span. That is, both sides share the same information

Generalization. The summary and document spans are event-coreferred[3]. The summary span does not explicitly mention specific details but instead uses broader terms that encompass those details. As a result, while the source span entails the summary span, the summary span does not fully entail the source span.

Conclusion. The summary span adds new information that is not mentioned explicitly in the source but derived from it. Accordingly, while the source span entails the summary span, the summary does not entail the source in full, and they are not event-coreferred.

The examples in Table 1 demonstrate how these guidelines are applied. More examples can be seen in the Appendix (Table 5).

[3]According to Eirew et al. (2022), event-coreferred spans are spans that refer to the same event.

Dataset	Paraph.	General.	Conc.	NA
DUC	65.3	13.8	11.1	9.8
FewSum	55.8	20.3	9.8	14.1
MultiNews	60.6	12.5	16.2	10.7

Table 2: Average distribution of span-level abstractiveness types in reference summaries (%)

Note that by definition, if a pair contains more than one type, it should be classified according to the more lenient type. For example, if one part of a summary span generalizes while other parts derive a conclusion, the spans are not considered *fully* coreferred and should be classified as a Conclusion. Thus, the final hierarchy is: Conclusion > Generalization > Paraphrase.

4 Annotation Process

In order to understand how often different levels of abstractiveness appear in human written summaries, we annotate reference summaries from the news and review domains. This annotation was performed manually by an expert annotator.

4.1 Alignment data

As outlined in Section 3, abstractiveness levels are determined by analyzing matching summary-source pairs. To achieve this, we utilize existing human-annotated source-summary span alignments from three multi-document summarization datasets: two from the news domain—DUC (NIST, 2014) and MultiNews (Fabbri et al., 2019)—and one focused on business reviews, FewSum (Bražinskas et al., 2020). Specifically, we used 315 document-summary span pairs across 12 summaries from DUC alignments (Ernst et al., 2021), 250 pairs from 16 summaries from MultiNews alignments (Ernst et al., 2024), and 336 pairs from 17 summaries using FewSum alignments (Slobodkin et al., 2024). For each pair, the annotator classified the abstractiveness type.

4.2 Annotation

The annotation process is composed by preliminary alignment data cleaning, pair-level annotation, aggregation to the span level, and finalizing rate calculation across all summaries. In this section, we elaborate about each of these components.

As we heavily rely on the previously annotated source-summary alignment, to ensure the quality of our annotations, we first cleaned the alignment data. This data allowed for some degree of noise, details that appeared in the summary or source span but not in the other span from the pair, such as *'two brothers'* in the conclusion example from Table 1. To meet our definition of *tight* alignment, we first omitted tokens from both the summary and source spans, ensuring no unaligned tokens remained, provided that such omissions did not contradict or significantly alter the span's meaning.

Then, we annotate each source-summary pair with one of the three types. If entailment could not be achieved even after omitting details, the pair was labeled as 'not aligned'.

Since a single summary span can be aligned with multiple document spans, and the level of abstractiveness is defined at the pair level, it is necessary to aggregate the pair-level decisions into a single summary span-level label. Specifically, each summary span was aligned separately with each of its corresponding document spans, and abstractiveness was annotated for every summary–document pair. After completing the pair-level annotations, we aggregated all annotations associated with the same summary span to derive a final abstractiveness label for that span.

Since we do not know which document span influenced the summary span the most, we adopt a strict approach, assuming that summarizers use the simplest level of abstractiveness possible. Specifically, if a summary span is aligned with multiple document spans, each annotated with a different abstractiveness type, we assume the summarizer employed the least abstract type. Paraphrase is the simplest type, followed by Generalization, and finally Conclusion. Therefore, a summary span is considered derived from Conclusion only if all associated pairs are labeled as Conclusion. If all pairs are labeled 'not aligned', the summary span is considered 'not aligned'.

Finally, we calculate the rate of each type in a summary and averaged across summaries in the same dataset.

4.3 Quality Evaluation

To assess the clarity of the task and the reliability of the guidelines, we recruited an additional expert annotator to independently annotate a subset of three summaries (one from each dataset) and measured inter-annotator agreement. Agreement was evaluated under two conditions: (1) for the full process, which included both span cleaning and classification; and (2) for classification only, where the annotator received the already cleaned spans. The annotators reached an agreement of 81.5% on the classification decisions and 69.2% on the full process when including the alignment-tightening process. These results indicate that the annotation guidelines are clear and can be applied consistently across annotators.

4.4 Results

As shown in Table 2, while Paraphrasing is the most common level of abstractiveness, approximately 25-35% of the reference summaries include instances of Generalization and Conclusion. This indicates that human summarizers frequently go beyond simple paraphrasing. Generalization is more prevalent than Conclusion in the reviews dataset, where summarizers often generalize across multiple personal experiences reported by customers. This tendency is more expected in reviews but less common in the news domain, where summarization typically focuses on factual reporting.

5 System-Generated Summaries

We observed a substantial presence of generalization and conclusion actions in human-written reference summaries, raising questions about whether automatic summarization systems are capable of generating similar types of information. However, due to the lack of available alignment data between system-generated summaries and their corresponding source documents, data required for our annotation process, we were unable to apply the same fine-grained annotation procedure to system summaries as we did for human references.

Instead, we analyzed how well system-generated summaries align with reference summaries across different abstractiveness levels. Our findings reveal that system outputs tend to match paraphrase-based reference spans far more frequently than generation- or conclusion-based spans. In other words, the high similarity scores that system summaries achieve relative to reference summaries primarily stem from their ability to produce effective paraphrases, rather than from generating novel conclusions or inferred content akin to those written by humans.

To examine this phenomenon, we conducted an abstractiveness-aware manual evaluation inspired by the Pyramid method (Nenkova and Passonneau, 2004). We selected 10 topics from MultiNews, 10 from FewSum, and 6 from DUC 2004. Sys-

Model	Paraph.	General.	Conc.	Total
GPT 4o m.	54.0	34.1	30.9	38.1
PRIMERA	35.1	16.7	0.0	20.8
Llama 3	55.6	35.5	33.1	39.1

Table 3: Average span-level recall scores of system summaries for each abstractiveness type

Reference	System
Democrats, meanwhile, argue that it's too soon to scale back the program	Democrats argue against these cuts, pointing out that many families still depend on the program
The quality of the theater is superb	the quality...justify the cost

Table 4: Examples of reference spans and related system spans with different abstractiveness levels.

tem summaries were generated using three models: PRIMERA fine-tuned on MultiNews (Xiao et al., 2022), GPT-4o mini in a zero-shot setting (OpenAI, 2024), and Llama 3 8B Instruct in a zero-shot setting (Dubey et al., 2024).

Following the standard Pyramid approach, an expert annotator segmented each reference summary into factual, standalone spans and then identified which of these were matched by the corresponding system-generated summary. The overall recall of matched spans for each model is reported in Table 3 ("Total").

Building on these matches, we computed a recall score for each abstractiveness level, reflecting the proportion of reference spans of a given type that were successfully reproduced by the system. As shown in Table 3, coverage varied substantially across types. For instance, GPT-4o mini matched 54% of reference spans classified as Paraphrase, but only 30.9% of those classified as Conclusion. Across all systems, Paraphrasing achieved the highest recall, while Conclusion and Generation spans were captured less frequently.

These findings suggest that current summarization models excel at reproducing paraphrased content but struggle to incorporate reasoning-based or conclusion-oriented information that match the reference summary, particularly when such reasoning is not explicitly required by the input.

It is important to note that, in a few cases, models did select the same source information as the reference summary did, but because they employed a different level of abstraction, their output was too distant from the reference to be considered a match. Examples of this phenomenon are provided in Table 4 and the Appendix (Table 6). This highlights that while models can identify relevant content, they may require further guidance on how to apply the appropriate level of abstraction.

6 Conclusion

In this work, we analyzed different abstraction levels in summarization, and found that while humans use reasoning to derive information which improves the focus and clarity of summaries, models are still lagging behind. We release our data and annotation to facilitate research in this direction and the development of summarization models that incorporate better reasoning abilities.

Limitations

Due to a lack of available system summary-source alignment data at the time of this project, we were unable to perform system summary annotation of the same kind of the reference summary annotation. From our analysis alone, we cannot conclude that system generated summaries do or do not contain generalization or conclusion information, only that they do not often apply these levels of abstractiveness in the same way as human summarizers.

Abstraction level of summary-source pairs were annotated, separately, one pair at a time. As a result, some summary spans may have been marked 'not aligned' if they require information fusion across many source spans.

Acknowledgments

We thank the anonymous reviewers for their constructive comments. This work was supported in part by the IVADO Postdoctoral Fellowship, Canada CIFAR AI Chair, and the Natural Sciences and Engineering Research Council of Canada.

References

Regina Barzilay and Kathleen R. McKeown. 2005. Sentence Fusion for Multidocument News Summarization. *Computational Linguistics*, 31(3):297–328.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Few-shot learning for opinion summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4119–4135, Online. Association for Computational Linguistics.

Danny Driess, F. Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Ho Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Peter R. Florence. 2023. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro

Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models.

Alon Eirew, Avi Caciularu, and Ido Dagan. 2022. Cross-document event coreference search: Task, dataset and modeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 900–913, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ori Ernst, Avi Caciularu, Ori Shapira, Ramakanth Pasunuru, Mohit Bansal, Jacob Goldberger, and Ido Dagan. 2022. Proposition-Level Clustering for Multi-Document Summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1765–1779, Seattle, United States. Association for Computational Linguistics.

Ori Ernst, Ori Shapira, Ramakanth Pasunuru, Michael Lepioshkin, Jacob Goldberger, Mohit Bansal, and Ido Dagan. 2021. Summary-source proposition-level alignment: Task, datasets and supervised baseline. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 310–322, Online. Association for Computational Linguistics.

Ori Ernst, Ori Shapira, Aviv Slobodkin, Sharon Adar, Mohit Bansal, Jacob Goldberger, Ran Levy, and Ido Dagan. 2024. The power of summary-source alignments. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6527–6548, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Jessica Nayeli López Espejel, El Hassane Ettifouri, Mahaman Sanoussi Yahaya Alassan, El Mehdi Chouham, and Walid Dahhane. 2023. Gpt-3.5, gpt-4, or bard? evaluating llms reasoning ability in zero-shot setting and performance boosting through prompts. *Nat. Lang. Process. J.*, 5:100032.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

José Ángel González, Annie Louis, and Jackie Chi Kit Cheung. 2022. Source-summary entity aggregation in abstractive summarization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6019–6034, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *ArXiv*, abs/2209.12356.

Jingyi He, Meng Cao, and Jackie Chi Kit Cheung. 2023. Analyzing multi-sentence aggregation in abstractive summarization via the shapley value. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 121–134, Singapore. Association for Computational Linguistics.

Hongyan Jing. 2002. Using hidden Markov modeling to decompose human-written summaries. *Computational Linguistics*, 28(4):527–543.

Clément Jumel, Annie Louis, and Jackie Chi Kit Cheung. 2020. TESA: A Task in Entity Semantic Aggregation for abstractive summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8031–8050, Online. Association for Computational Linguistics.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Logan Lebanoff, John Muchovej, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019a. Analyzing sentence fusion in abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 104–110, Hong Kong, China. Association for Computational Linguistics.

Logan Lebanoff, John Muchovej, Franck Dernoncourt, Doo Soon Kim, Lidan Wang, Walter Chang, and Fei Liu. 2020. Understanding points of correspondence between sentences for abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 191–198, Online. Association for Computational Linguistics.

Logan Lebanoff, Kaiqiang Song, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019b. Scoring Sentence Singletons and Pairs for Abstractive Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2175–2189, Florence, Italy. Association for Computational Linguistics.

Gabrielle Kaili-May Liu, Bowen Shi, Avi Caciularu, Idan Szpektor, and Arman Cohan. 2025. MDCure: A scalable pipeline for multi-document instruction-following. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29258–29296, Vienna, Austria. Association for Computational Linguistics.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating Content Selection in Summarization: The Pyramid Method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

NIST. 2014. Document understanding conferences. `https://duc.nist.gov`. Accessed: 2019-12-02.

OpenAI. 2024. [link].

David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. 2019. Analysing mathematical reasoning abilities of neural models. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. 2019. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4506–4515, Hong Kong, China. Association for Computational Linguistics.

Aviv Slobodkin, Ori Shapira, Ran Levy, and Ido Dagan. 2024. Multi-review fusion-in-context. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3003–3021, Mexico City, Mexico. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Ming Tu, Guangtao Wang, Jing Huang, Yun Tang, Xiaodong He, and Bowen Zhou. 2019. Multi-hop reading comprehension across multiple documents by reasoning over heterogeneous graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2704–2713, Florence, Italy. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. 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)*, pages 5245–5263, Dublin, Ireland. Association for Computational Linguistics.

Xingxing Zhang, Mirella Lapata, Furu Wei, and Ming Zhou. 2018. Neural Latent Extractive Document Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 779–784, Brussels, Belgium. Association for Computational Linguistics.

Wenting Zhao, Justin Chiu, Claire Cardie, and Alexander Rush. 2023. Hop, union, generate: Explainable multi-hop reasoning without rationale supervision. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages

16119–16130, Singapore. Association for Computational Linguistics.

A Reference Summary Examples

Table 5 contains examples of aligned pairs from the datasets and their corresponding types after the annotation process.

B System Summary Examples

In Table 6, we provide examples of similar spans from the reference and system summaries. Summaries generated with GPT 4o mini and Llama 3 used the prompt, "The following are news articles on a single topic. Please create one summary from all of the articles:", followed the source documents.

C System Summary Evaluation by Dataset

We present the system summary evaluation statistics broken down by dataset in Table 7.

Source	Summary	Type
People now can look at Hawaii as a destination to have their marriage	That could be a boon for destination weddings in a state	Conclusion
a source tells Us that Maroon 5 have been tapped to grace the halftime stage	sources tell Variety the Super Bowl LIII halftime act has been chosen, and it's Maroon 5	Paraphrase
the court will return to the subject of whether the Constitution permits public colleges and universities to take account of race in admissions decisions	The court...will look at cases including...affirmative action	Generalization
an estimated 16 million children, or about one in five, received food stamp assistance	about 16 million, or nearly one in five, of them are doing so fueled by food stamps	Paraphrase
"I actually quite like the color," said psychologist Dr. Carolyn Mair..."It's an earthy, muted, rich color, very much of nature	But not everyone's on board with trashing opaque couche	Generalization
You should always use the chip device, not the swipe device	Credit card purchases are about to get...a lot more secure	Not Aligned
Can the sight of a greenish-brown color really be enough to deter smokers from reaching for their next pack of cigarettes?	the...color... been used to try to save lives	Conclusion
waited over 20 min	The food takes to long to come out	Generalization
waited over 20 min ... not even a sorry about the wait	service is horrible	Conclusion
newest hottest spot ... Food is amazing	Overall not a recommended place	Not Aligned

Table 5: Examples of reference spans and related system spans with different abstractiveness levels.

Reference Type	Reference	System	Matched
Generalization	Instagram...simply dropping the changes	Instagram has decided to revert to its previous terms	Yes
Paraphrase	The chips in the new cards use a system known as EMV, for creators Europay, MasterCard, and Visa	These EMV (Europay, MasterCard, and Visa) chips	Yes
Conclusion	police "do not believe there are any outstanding suspects," per a spokesperson	authorities are not currently seeking any suspects	No
Conclusion	Flickr's app jumped in popularity	NA	No
Generalization	the opposition tried to cut off his access to loans.	opposition...urged international bodies, such as the Asian Development Bank, to reconsider support	No
Generalization	Police got a call...that afternoon	Police...receiving a distress call around 2:45 p.m	Yes
Paraphrase	Crowe eventually returned to the matter of the accent, saying, "I'm a little dumbfounded you could possibly find any Irish in that character-that's kind of ridiculous, but it's your show.	Crowe became defensive and denied the accusation	No
Conclusion	the investigation is ongoing	The investigation is ongoing	Yes

Table 6: Examples of reference spans and related system spans with different abstractiveness levels.

Model	Dataset	Paraphrase	Generalization	Conclusion	Total
GPT 4o mini	DUC2004	44.7	8.3	66.7	34.8
	FewSum	63.5	37.5	17.9	41.9
	MultiNews	51.0	41.6	29.1	36.3
PRIMERA	DUC2004	23.1	0.0	0.0	16.5
	FewSum	34.8	13.3	0.0	16.5
	MultiNews	42.7	27.8	0.0	27.7
Llama 3 8b	DUC2004	44.7	33.3	0.0	37.7
	FewSum	72.6	33.4	35.7	42.6
	MultiNews	46.9	38.9	42.0	36.4

Table 7: Average span-level recall scores of system summaries for each abstractiveness type and dataset.