

While the CoD prompting strategy is interesting—and based on data/examples provided, seems to work—I would benefit from a comparison of different approaches that were taken to generate the final prompt presented in this paper. Maybe CoD could be more effective if the LLM itself didn't implement the 5-turns? What happens if CoD is run for 100 turns?

From Sparse to Dense: GPT-4 Summarization with Chain of Density Prompting

Griffin Adams♣♣
griffin.adams@columbia.edu

Alexander R. Fabbri◇
afabbri@salesforce.com

Faisal Ladhak ♠
faisal@cs.columbia.edu

Eric Lehman♥
lehmer16@mit.edu

Noémie Elhadad♣♣
noemie.elhadad@columbia.edu

Salesforce AI◇ MIT♥ Columbia University: CS♣, Biomedical Informatics♣

Abstract

Selecting the “right” amount of information to include in a summary is a difficult task. A good summary should be detailed and entity-centric without being overly dense and hard to follow. To better understand this tradeoff, we solicit increasingly dense GPT-4 summaries with what we refer to as a “Chain of Density” (CoD) prompt. Specifically, GPT-4 generates an initial entity-sparse summary before iteratively incorporating missing salient entities without increasing the length. Summaries generated by CoD are more abstractive, exhibit more fusion, and have less of a lead bias than GPT-4 summaries generated by a vanilla prompt. We conduct a human preference study on 100 CNN DailyMail articles and find that humans prefer GPT-4 summaries that are more dense than those generated by a vanilla prompt and almost as dense as human written summaries. Qualitative analysis supports the notion that there exists a tradeoff between informativeness and readability. 500 annotated CoD summaries, as well as an extra 5,000 unannotated summaries, are freely available on HuggingFace¹.

1 Introduction

Automatic summarization has come a long way in the past few years, largely due to a paradigm shift away from supervised fine-tuning on labeled datasets to zero-shot prompting with Large Language Models (LLMs), such as GPT-4 (OpenAI, 2023). Without additional training, careful prompting can enable fine-grained control over summary characteristics, such as length (Goyal et al., 2022), topics (Bhaskar et al., 2023), and style (Pu and Demberg, 2023).

An overlooked aspect is the information density of an summary. In theory, as a compression of another text, a summary *should* be denser—containing a higher concentration of information—than the source document. Given the high latency of LLM decoding (Kaddour et al., 2023), covering more information in fewer

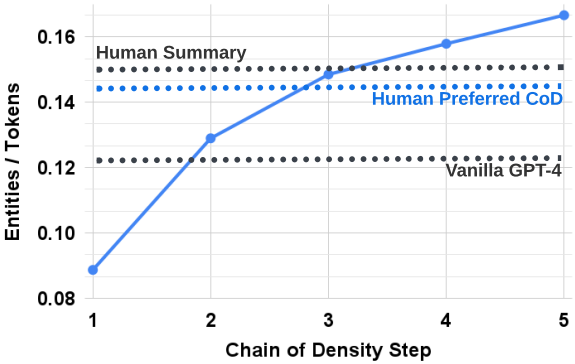


Figure 1: Chain of Density (CoD) summaries grow increasingly entity dense, starting off closer to vanilla GPT-4 summaries and eventually surpassing that of human written summaries. Human annotations suggest that a density similar to that of human-written summaries is preferable—striking the right balance between clarity (favors less dense) and informativeness (favors more dense).

words is a worthy goal, especially for real-time applications. Yet, how dense is an open question. A summary is uninformative if it contains insufficient detail. If it contains too much information, however, it can become difficult to follow without having to increase the overall length. Conveying more information subject to a fixed token budget requires a combination of abstraction, compression, and fusion. There is a limit to how much space can be made for additional information before becoming illegible or even factually incorrect.

In this paper, we seek to identify this limit by soliciting human preferences on a set of increasingly dense summaries produced by GPT-4. Treating entities, and, in particular, the average number of entities per token, this would as a proxy for density, we generate an initial, entity-sparse summary. Then, we iteratively identify and fuse a value with need to be a value with ≤ 1 , ya? 1-3 missing entities from the previous summary without increasing the overall length (5x overall). Each summary has a higher ratio of entities to tokens than the previous one. Based on human preference data, we determine that humans prefer summaries that are almost as dense as human-written summaries and more

CoD is a compression algorithm using LLMs

• Could you test CoD by comparing performance of raw prompts vs. summarized or “compressed” prompts for a variety of task types?

¹https://huggingface.co/datasets/griffin/chain_of_density

this is asking the LLM to perform a for loop. I wonder if it would be more effective to ask the LLM to do 1 turn at a time and present the last summary only instead of it producing all 5 summaries at once.

Chain of Density (CoD) Prompt

```
Article: {{ARTICLE}}
You will generate increasingly concise, entity-dense summaries of the
above Article.
Repeat the following 2 steps 5 times.
Step 1. Identify 1-3 informative Entities (";" delimited) from the
Article which are missing from the previously generated summary.
Step 2. Write a new, denser summary of identical length which covers
every entity and detail from the previous summary plus the Missing
Entities.
A Missing Entity is:
- Relevant: to the main story.
- Specific: descriptive yet concise (5 words or fewer).
- Novel: not in the previous summary.
- Faithful: present in the Article.
- Anywhere: located anywhere in the Article.
Guidelines:
- The first summary should be long (4-5 sentences, ~80 words) yet
highly non-specific, containing little information beyond the
entities marked as missing. Use overly verbose language and fillers
(e.g., "this article discusses") to reach ~80 words.
- Make every word count: re-write the previous summary to improve
flow and make space for additional entities.
- Make space with fusion, compression, and removal of uninformative
phrases like "the article discusses".
- The summaries should become highly dense and concise yet
self-contained, e.g., easily understood without the Article.
- Missing entities can appear anywhere in the new summary.
- Never drop entities from the previous summary. If space cannot be
made, add fewer new entities.
Remember, use the exact same number of words for each summary.
Answer in JSON. The JSON should be a list (length 5) of dictionaries
whose keys are "Missing_Entities" and "Denser_Summary".
```

CoD Outputs (Added Details from previous)

This article discusses an incident that occurred during the Chinese Grand Prix involving two racing drivers, Jenson Button and Pastor Maldonado. The two were competing for the 13th place when Button collided with Maldonado's vehicle, causing damage to both cars. The incident resulted in a penalty for Button, who was demoted to 14th place. Maldonado, on the other hand, had to retire from the race due to the damage his car sustained.

During the Chinese Grand Prix, Jenson Button, driving for McLaren, collided with Pastor Maldonado's Lotus, while they were vying for 13th place. The crash caused Maldonado to retire and Button to receive a penalty, dropping him to 14th. The incident unfolded in front of Fernando Alonso, who managed to avoid the collision and move up two places.

In the Chinese Grand Prix, Jenson Button's McLaren hit Pastor Maldonado's Lotus, causing Maldonado to retire and Button to receive a five-second penalty, demoting him to 14th. Button also received two penalty points on his superlicence. Fernando Alonso, who witnessed the incident, advanced two places, while Button was lapped by Nico Rosberg's Mercedes.

Jenson Button's McLaren collided with Pastor Maldonado's Lotus during the Chinese Grand Prix, causing front wing damage to Button's car and rear-end damage to Maldonado's, forcing his retirement. Button received a five-second penalty and two superlicence points, dropping him to 14th. Fernando Alonso advanced two places, while Button was lapped by Nico Rosberg and Alonso by Sebastian Vettel and Kimi Raikkonen.

On lap 49 of the incident-packed Chinese Grand Prix, Jenson Button's McLaren hit Pastor Maldonado's Lotus, causing damage to Maldonado's retirement. Button received a five-second penalty and two superlicence points, falling to 14th. Fernando Alonso, who witnessed the crash, advanced two places, while Button was lapped by Nico Rosberg and Alonso by Ferrari's Sebastian Vettel and Kimi Raikkonen.

here we
lose information
about the
type of
damage

Figure 2: Chain of Density (CoD) Prompt and example output. At each step, 1-3 additional details (entities) are added to the previous summary without increasing the length. To make room for new entities, existing content is re-written (e.g., compression, fusion). Half the annotators (2/4) prefer the second to last summary, with the others preferring the final one.

Small Sample Size

dense than those generated by a vanilla GPT-4 prompt. Our primary contributions are to:

- Develop a prompt-based iterative method (CoD) for making summaries increasingly entity dense.
- Conduct both human and automatic evaluation of increasingly dense summaries on CNN/DailyMail articles to better understand the tradeoff between informativeness (favoring more entities) and clarity (favoring fewer entities).
- Open source GPT-4 summaries, annotations, and a set of 5,000 unannotated CoD summaries to be used for evaluation or distillation.

To maintain the same length while increasing the number of entities covered, abstraction, fusion, and compression is explicitly encouraged, rather than dropping meaningful content from previous summaries.

Figure 2 displays the prompt along with an example output. Rather than be prescriptive about the types of entities, we simply define a Missing Entity as:

- **Relevant:** to the main story.
- **Specific:** descriptive yet concise (5 words or fewer).
- **Novel:** not in the previous summary.
- **Faithful:** present in the Article.
- **Anywhere:** located anywhere in the Article.

Data. We randomly sample 100 articles from the CNN/DailyMail summarization (Nallapati et al., 2016) test set for which to generate CoD summaries.

Reference Points. For frame of reference, we compare CoD summary statistics to human-written bullet-point style reference summaries as well as summaries generated by GPT-4 with a vanilla prompt: "Write a VERY short summary of the Article. Do not exceed 70 words." We set the desired token length to match that of CoD summaries (shown in Table 1).

3 Statistics

Direct statistics (tokens, entities, entity density) are ones directly controlled for by CoD, while Indirect

2 Chain of Density Prompting

Prompt. Our goal is to generate a set of summaries with GPT-4 with varying levels of information density, while controlling for length, which has proven to be a strong confounder when evaluating summaries (Fabbri et al., 2021; Liu et al., 2023b). To do this, we formulate a single Chain of Density (CoD) prompt, whereby an initial summary is generated and made increasingly entity dense. Specifically, for a fixed number of turns, a set of unique salient entities from the source text are identified and fused into the previous summary without increasing the length. The first summary is entity-sparse as it focuses on only 1-3 initial entities.

how was it determined that the original summary from gpt4 only has 1-3 entities regardless of the initial content?

what is the point of diminishing return for this algorithm where adding "density" actually "dilutes" the more important meaning of the text?
is the impact of this measured?

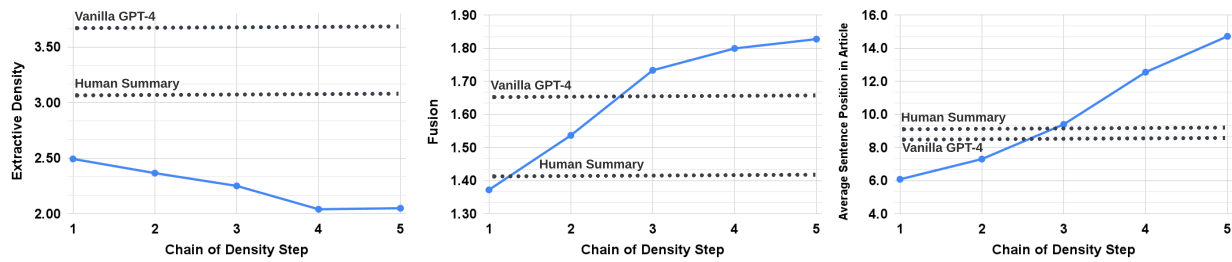


Figure 3: **CoD**-generated summaries grow increasingly abstractive while exhibiting more fusion and less of a lead bias.

statistics are expected byproducts of densification.

CoD Step	Tokens	Entities	Density (E/T)
1	72	6.4	0.089
2	67	8.7	0.129
3	67	9.9	0.148
4	69	10.8	0.158
5	72	12.1	0.167
Human	60	8.8	0.151
Vanilla GPT-4	70	8.5	0.122

Table 1: Explicit statistics for GPT-4 **CoD** summaries.

Direct Statistics. In Table 1, we compute tokens with NLTK (Loper and Bird, 2002), measure unique entities with Spacy², and compute entity density as the ratio. The **CoD** prompt largely adheres to a fixed token budget. In fact, the second step leads to an average 5-token (72 to 67) reduction in length as unnecessary words are removed from the initially verbose summary. The entity density rises—starting at 0.089, initially below Human and Vanilla GPT-4 (0.151 and 0.122)—to 0.167 after 5 steps of densification.

Indirect Statistics. *Abstractiveness* should increase with each **CoD** step because summaries are iteratively re-written to make space for each additional entity. We measure abstractiveness with extractive density: the average squared length of extractive fragments (Grusky et al., 2018). Similarly, the level of concept *Fusion* should increase monotonically as entities are added to a fixed-length summary. We proxy fusion as average number of source sentences aligned to each summary sentence. For alignment, we use the relative ROUGE gain method (Zhou et al., 2018), which aligns source sentences to a target sentence until the relative ROUGE gain of an additional sentence is no longer positive. We also expect the *Content Distribution*—the position in the Article from which summary content is sourced—to shift. Specifically, we expect that **CoD** summaries initially exhibit a strong Lead Bias yet gradually start to pull in entities from the

middle and end of the article. To measure this, we use our alignments from fusion and measure the average sentence rank of all aligned source sentences. Figure 3 confirms these hypotheses: abstractiveness increases with the number of re-writing steps (lower extractive density on the left), the rate of fusion rises (middle figure), and the summaries start to incorporate content from the middle and end of the article (right figure). Interestingly, all **CoD** summaries are more abstractive than both human written and baseline summaries.

4 Results

To better understand the tradeoffs present with **CoD** summaries, we conduct a preference-based human study and a rating-based evaluation with GPT-4.

CoD Step	% Share of First Place Votes				
	Individual Annotators				Aggregate
1	3.0	2.0	13.0	17.4	8.3
2	25.0	28.0	43.0	31.4	30.8
3	22.0	28.0	21.0	24.4	23.0
4	29.0	25.0	13.0	26.7	22.5
5	21.0	17.0	10.0	16.3	15.5

Table 2: Breakdown of first-place votes for **CoD** summaries by step. Based on aggregate preferences, the modal **CoD** step is 2, median is 3, and expected is 3.06.

Human Preferences. We conduct a human evaluation to assess the impact of densification on human assessments of overall quality. Specifically, the first four authors of the paper were presented with randomly shuffled **CoD** summaries, along with the articles, for the same 100 articles (5 steps * 100 = 500 total summaries). Based on the definition of a “good summary” from Stiennon et al. (2020) (Table 6 from their paper), each annotator indicated their top preferred summary. Table 2 reports the breakdown of first place votes by **CoD** step across annotators—as well as aggregated across annotators. First, we report a low Fleiss’ kappa (Fleiss, 1971) of 0.112, which points to the subtle differences between summaries and the subjective nature of the task. Recent work has

²<https://spacy.io>.

CoD Step	Entity Density	Informative	Quality	Coherence	Attributable	Overall	GPT-4 Eval Average
1	0.089	4.34	4.75	4.96	4.96	4.41	4.69
2	0.129	4.62	4.79	4.92	5.00	4.58	4.78
3	0.148	4.67	4.76	4.84	5.00	4.57	4.77
4	0.158	4.74	4.69	4.75	5.00	4.61	4.76
5	0.167	4.73	4.65	4.61	4.97	4.58	4.71

Table 3: GPT-4 Likert-scale (1-5) assessments of Chain of Density (CoD) Summaries by step.

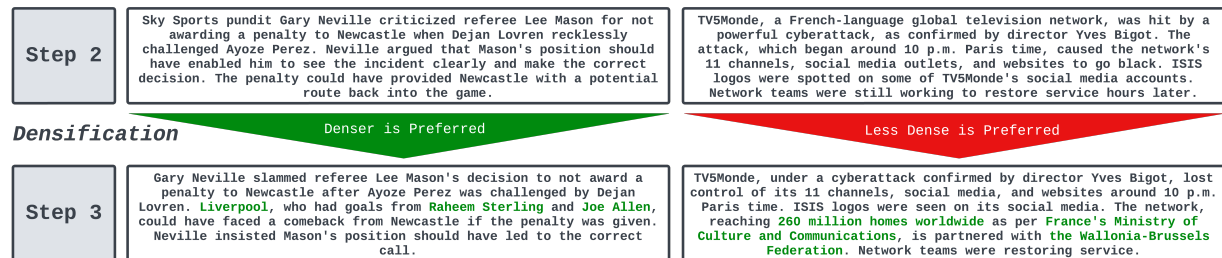


Figure 4: An example of a human-preferred densification step (left) and one which is not preferred. For the left, the bottom summary is preferred because the addition of “Liverpool” and the goal-scorers is relevant. The second summary makes room with sensible compressions, such as synthesizing “a potential route back into the game” into “a comeback”. For the right, the addition of more details on “TVMonde” does not make up for the presence of an awkward fusion of entities (“cyberattack”, and “Yves Bigot”), which was a direct result of having to tighten the previous summary.

similarly noted low instance-level agreement when judging GPT-based summaries (Goyal et al., 2022).

Yet, at the system level, some trends start to emerge. For 3 of the 4 annotators, CoD step 1 received the largest share of first-place votes across the 100 examples (28, 43, and 31.4%, respectively). Yet, in aggregate, 61% of first placed summaries (23.0+22.5+15.5) involved ≥ 3 densification steps. The median preferred CoD step is in the middle (3), and the expected step is 3.06.

Based on the average density of Step 3 summaries, we can roughly infer a preferred entity density of ~ 0.15 across the CoD candidates. From Table 1, we can see that this density aligns with human-written summaries (0.151), yet is noticeable higher than summaries produced with a vanilla GPT-4 prompt (0.122).

Automatic Metrics. As an evaluator, GPT-4 has been shown to adequately correlate to human judgments (Fu et al., 2023; Liu et al., 2023a), even potentially outperforming crowd-sourced workers on some annotation tasks (Gilardi et al., 2023). As a complement to our human evaluation (below), we prompt GPT-4 to rate CoD summaries (1-5) along 5 dimensions: **Informative**, **Quality**, **Coherence**, **Attributable**, and **Overall**. The definitions of **Informative**, **Quality**, and **Attributable** come from Aharoni et al. (2023), while **Coherence** comes from Fabbri et al. (2021)³. **Overall** aims to capture the qualities jointly. Please see Appendix A for the prompts used

³Quality and Coherence are article-independent metrics.

to solicit scores for each dimension. Table 3 suggests that densification is correlated with informativeness, yet there is a limit, with the score peaking at Step 4 (4.74). Article-free dimensions: **Quality** and **Coherence**, decline sooner (after 2 and 1 steps, respectively). All summaries are deemed **Attributable** to the source article. The **Overall** scores skew toward denser and more informative summaries, with **Step 4** having the highest score. On average across dimensions, the first and last CoD steps are *least* favored, while the middle three are close (4.78, 4.77, and 4.76, respectively).

In Appendix A, we report highest summary-level correlations of the **Overall** metric to human judgments (0.31 Pearson correlation), yet note low correlations overall—a phenomenon observed by Deutsch et al. (2022) when summaries are of similar quality.

Qualitative Analysis. There exists a clear trade-off between coherence / readability of summaries and informativeness. To illustrate, in Figure 4, we present two CoD steps: one for which the summary is improved with more detail, and one for which the summary is harmed. On average, intermediate CoD summaries best achieved this balance, yet we leave it to future work to precisely define and quantify this tradeoff.

5 Related Work

GPT Summarization. Goyal et al. (2022) benchmarked GPT-3 on news article summarization and found that humans preferred GPT-3 summaries over previous supervised baselines, which was

I think the authors should have motivated this evaluation structure more. For example — does inclusion or exclusion of “overall” correlate more to human preference.

how is it 'harmed'?

not reflective of existing reference-based and reference-free metrics. Zhang et al. (2023) find that zeroshot GPT-3 summaries perform on par with humans by soliciting high-quality summaries from freelance writers. **Entity-Based Summarization.** Narayan et al. (2021) proposed generating entity chains as a planning step for supervised fine-tuning of summarization models, in contrast to keywords (Li et al., 2020; Dou et al., 2021) or purely extractive units (Dou et al., 2021; Adams et al., 2023a). Entities have also been incorporated for summarization as a form of control (Liu and Chen, 2021; He et al., 2022; Maddela et al., 2022), to improve faithfulness (Nan et al., 2021; Adams et al., 2022), and as a unit for evaluation (Cao et al., 2022; Adams et al., 2023b).

6 Conclusion

We study the impact of summary densification on human preferences of overall quality. We find that a degree of densification is preferred, yet, when summaries contain too many entities per token, it is very difficult maintain readability and coherence. We open-source annotated test set as well as a larger un-annotated training set for further research into the topic of fixed-length, variable density summarization.

7 Limitations

We only analyze **CoD** for a single domain, news summarization. Annotations did not show high summary-level agreement yet did start to show system-level trends, which is in line with previous work on LLM-based evaluation (Goyal et al., 2022). Finally, GPT-4 is a closed source model so we cannot share model weights. We do, however, publish all evaluation data, annotations, as well as 5,000 un-annotated **CoD** to be used for downstream uses cases, e.g., density distillation into an open-sourced model such as LLAMA-2 (Touvron et al., 2023).

References

- Griffin Adams, Alex Fabbri, Faisal Ladhak, Noémie Elhadad, and Kathleen McKeown. 2023a. **Generating EDU extracts for plan-guided summary re-ranking.** In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2680–2697, Toronto, Canada. Association for Computational Linguistics.
- Griffin Adams, Han-Chin Shing, Qing Sun, Christopher Winestock, Kathleen McKeown, and Noémie Elhadad. 2022. **Learning to revise references for faithful summarization.** In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4009–4027, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Griffin Adams, Jason Zucker, and Noémie Elhadad. 2023b. **A meta-evaluation of faithfulness metrics for long-form hospital-course summarization.** *arXiv preprint arXiv:2303.03948*.
- Roei Aharoni, Shashi Narayan, Joshua Maynez, Jonathan Herzig, Elizabeth Clark, and Mirella Lapata. 2023. **Multilingual summarization with factual consistency evaluation.** In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3562–3591, Toronto, Canada. Association for Computational Linguistics.
- Adithya Bhaskar, Alex Fabbri, and Greg Durrett. 2023. **Prompted opinion summarization with GPT-3.5.** In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9282–9300, Toronto, Canada. Association for Computational Linguistics.
- Meng Cao, Yue Dong, and Jackie Cheung. 2022. **Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization.** In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.
- Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. **Re-examining system-level correlations of automatic summarization evaluation metrics.** In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6038–6052, Seattle, United States. Association for Computational Linguistics.
- Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. **GSum: A general framework for guided neural abstractive summarization.** In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4830–4842, Online. Association for Computational Linguistics.
- Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. **SummEval: Re-evaluating summarization evaluation.** *Transactions of the Association for Computational Linguistics*, 9:391–409.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. **Gptscore: Evaluate as you desire.** *arXiv preprint arXiv:2302.04166*.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. **Chatgpt outperforms crowd-workers for text-annotation tasks.** *arXiv preprint arXiv:2303.15056*.
- Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. **News summarization and evaluation in the era of gpt-3.** *arXiv preprint arXiv:2209.12356*.

- Max Grusky, Mor Naaman, and Yoav Artzi. 2018. [Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.
- Junxian He, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2022. [CTRLsum: Towards generic controllable text summarization](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5879–5915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.
- Haoran Li, Junnan Zhu, Jiajun Zhang, Chengqing Zong, and Xiaodong He. 2020. Keywords-guided abstractive sentence summarization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8196–8203.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.
- Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023b. [Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.
- Zhengyuan Liu and Nancy Chen. 2021. [Controllable neural dialogue summarization with personal named entity planning](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 92–106, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.
- Mounica Maddela, Mayank Kulkarni, and Daniel Preotiuc-Pietro. 2022. [EntSUM: A data set for entity-centric extractive summarization](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3355–3366, Dublin, Ireland. Association for Computational Linguistics.
- Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. [Abstractive text summarization using sequence-to-sequence RNNs and beyond](#). In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.
- Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. [Entity-level factual consistency of abstractive text summarization](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2727–2733, Online. Association for Computational Linguistics.
- Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. [Planning with learned entity prompts for abstractive summarization](#). *Transactions of the Association for Computational Linguistics*, 9:1475–1492.
- OpenAI. 2023. [Gpt-4 technical report](#). *ArXiv*, abs/2303.08774.
- Dongqi Pu and Vera Demberg. 2023. [ChatGPT vs human-authored text: Insights into controllable text summarization and sentence style transfer](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1–18, Toronto, Canada. Association for Computational Linguistics.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2023. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848*.
- Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. [Neural document summarization by jointly learning to score and select sentences](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–663, Melbourne, Australia. Association for Computational Linguistics.

A GPT-4 Metrics

For the GPT-4 Likert-style evaluation, we use the following prompt template.

Article: {{Article}}

Summary: {{Summary}}

Please rate the summary
(1=worst to 5=best) with
respect to {{Dimension}}.

{{Definition}}

Below, we present the definitions provided for each quality metric.

- **Informative:** An informative summary captures the important information in the article and presents it accurately and concisely.
- **Quality:** A high quality summary is comprehensible and understandable.
- **Coherence:** A coherent summary is well-structured and well-organized.
- **Attributable:** Is all the information in the summary fully attributable to the Article?
- **Overall Preference:** A good summary should convey the main ideas in the Article in a concise, logical, and coherent fashion.

The **Quality** and **Coherence** prompts do not include the Article in the prompt. These definitions were paraphrased from previous summarization annotation efforts: (Fabbri et al., 2021; Aharoni et al., 2023).

Dimension	Correlation
Informative	0.215
Quality	0.120
Coherence	0.178
Attributable	0.245
Overall	0.311

Table 4: Summary-Level Pearson Correlation coefficient between human preferences and GPT-4 Likert ratings.

Meta-Evaluation. To compute the summary-level correlation, we first turned the preference data into a vector representing the number of times that summary received a first-placed vote. Table 4 demonstrates, unsurprisingly, that a prompt designed to capture overall summary rating has the highest summary-level Pearson correlation to overall preferences (31), yet overall correlations are still low.