GRASUM at BioLaySumm Task 1: Background Knowledge Grounding for Readable, Relevant, and Factual Biomedical Lay Summaries

Domenic Rosati
scite.ai / Brooklyn, N.Y.
Dalhousie University / Halifax N.S.

Abstract

Communication of scientific findings to the public is important for keeping non-experts informed of developments such as life-saving medical treatments. However, generating readable lay summaries from scientific documents is challenging, and currently, these summaries suffer from critical factual errors. One popular intervention for improving factuality is using additional external knowledge to provide factual grounding. However, it is unclear how these grounding sources should be retrieved, selected, or integrated, and how supplementary grounding documents might affect the readability or relevance of the generated summaries. We develop a simple method for selecting grounding sources and integrating them with source documents. We then use the BioLaySumm summarization dataset to evaluate the effects of different grounding sources on summary quality. We found that grounding source documents improves the relevance and readability of lay summaries but does not improve factuality of lay summaries. This continues to be true in zero-shot summarization settings where we hypothesized that grounding might be even more important for factual lay summaries [1].

1 Introduction

Automatic lay summarization of biomedical research is a promising approach to help inform non-experts of vital scientific and clinical discoveries. However, known issues with factuality in automatic summarization systems (Gabriel et al., 2021; Maynez et al., 2020) are still a barrier that prevents their safe deployment. To improve factuality, some have suggested using grounding sources with retrieval augmentation (Shuster et al., 2021; Lewis et al., 2021; Thoppilan et al., 2022). This has been show to help maintain factuality in biomedical nlp tasks (Guo et al., 2022) without harming

---

Lay Summary: [Messenger RNAs carry the instructions necessary to synthesize proteins that do work for the cell]$_{background}$ In this work , we surveyed mRNA ends from 10 , 000 genes in immune cells from genetically distinct human individuals.
Abstracts: Virtually all messenger RNAs (mRNAs) in eukaryotes are cleaved and polyadenylated at their 3 ends. **UMLS:** RNA is unique among biological macromolecules in that it can encode genetic information. **Wiki Simple:** Messenger RNA carries a genetic message from the DNA to the protein making machinery of the cell. **Wikipedia**: An mRNA molecule is transcribed from the DNA sequence and is later translated into protein.

Figure 1: Example of different types of background knowledge from grounding sources.

readability or relevance. Additionally, (Guo et al., 2022) suggests that retrieval augmentation is especially helpful for lay summarization because those summaries need to provide necessary background knowledge (see figure 1) such as definitions which are not often found in the source text.

In this paper we wanted to understand (i) how we might develop a retrieval augmentation solution for lay summarization when using whole scientific papers for models with limited input context length and (ii) what is the effect of different grounding sources that contain different types of background knowledge on readability, relevancy, and factuality.

Contributions: We develop (i) *a simple method for selecting and using grounding sources for lay summarization*. We assess this method using the BioLaySumm (Goldsack et al., 2022) biomedical paper lay summarization dataset[2] and find that (ii) *grounding has the largest effect on readability* where in the zero-shot summarization setting definitional background knowledge from Unified Medical Language System (UMLS) and simplified encyclopedic background knowledge from Wikipedia Simple provide better readability scores. Contrary to popular opinion, we found that (iii) *grounding*

---

[1] https://github.com/domenicrosati/improving-lay-factuality-with-retreival

[2] This paper participates in the BioLaySumm shared task 1 outlined in Goldsack et al. (2023)

does not improve factuality.

2 Method

Summarizing scientific papers requires more input tokens than a large language model can typically support due to memory constraints. The average token count for articles in BioLaySumm are 8,963 tokens for PLOS and 13,942 tokens for eLife with articles up to 45,563 tokens. Since we want to explore the effect of grounding articles with additional retrieved sources there is an even greater need for large token input support. Because of these factors, the base model we use in our experiments is the Longformer Encoder-Decoder (LED) (Beltagy et al., 2020) which supports an input token length of 16,384 tokens (see Appendix B for additional training and inference details).

Our method was designed to test the effect of different grounding sources on downstream summarization quality. In addition to definitional background knowledge from UMLS and encyclopedic background knowledge from Wikipedia which were used in Guo et al. (2022), we introduced two other retrieval sources, *Wikipedia Simple* for access to encyclopedic background knowledge in simpler terms and *Scientific Abstracts* for access to further contextual background knowledge that might emulate the additional supplementary knowledge an expert has when crafting a lay summary. In all, we tested the following four grounding sources (1) *UMLS* named entity definitions (2) *Scientific Abstracts* (from Crossref) (3) *Wikipedia* (English) (4) *Wikipedia Simple* (English). See Appendix C for a full description of these grounding sources and how they were used.

Our retrieval augmentation consisted of two steps: (i) retrieving and (ii) re-ranking documents, First we took each sentence in the leading 1,024 tokens of the article (roughly corresponding to the abstract) and searched them using BM25 on indexes constructed for each grounding source (except *UMLS* which uses another method discussed). Indexes were constructed using Pyserini (Lin et al., 2021). The top 1 most relevant passage is selected and then added to a pool of candidate passages. In the case of *Scientific Abstracts*, we remove the abstract of the document we are enhancing if it was in the pool. In the case of *UMLS*, we follow Guo et al. (2022) by using the scispaCy entity linker (Neumann et al., 2019) over the first 1,024 tokens and provide definitions for the UMLS named entities

as the pool of candidate passages.

The above procedure results in too many results to fit within context length. In order to resolve this, we rank the pool of candidates passages against the first 1,024 tokens of the using a crossencoder (Reimers and Gurevych, 2019) (see Appendix B for details). Finally we construct our inputs by selecting the first 8,192 tokens of the original document and the top relevant grounding passages up to 8,192 tokens. A <|SEARCH|> token is inserted between the original document and the grounding passage and global attention is placed on the <|SEARCH|> token in order to assist with attention over the grounding passages.

We also supplement all grounding sources with a bibliographic reference string containing the title, authors, and year of the paper being summarised. This was motivated by seeing many ground truth summaries which cited the source document by first authors name (For example: "Parks et al. analyzed data on US deaths between 1980 and 2016" which is the first sample in the eLife training subset of BioLaySumm).

3 Experiments

3.1 BioLaySumm

We experiment with the method above using the BioLaySumm lay summarization dataset which contains 29,119 training samples (24,773 from PLOS and 4,346 from eLife) and 1,617 validation samples (1,376 from PLOS and 241 from eLife). See Appendix A for more details. This dataset is used to evaluate the ability for models to provide factual, readable, and relevant lay summaries of biomedical research articles from research papers in PLOS and eLife which are paired with human written lay summaries. All of the methods use a LED base model trained for 4 epochs evaluated every 5,000 steps, like (Goldsack et al., 2022) the checkpoint with the best ROUGE-2 is selected (see Appendix B for more details).

Relevancy was measured based on BERTScore, Rouge1, Rouge2, and RougeL. Factuality was measured using a BARTScore (Yuan et al., 2021) trained on the BioLaySumm dataset as well as an unsupervised metric SummaC (Laban et al., 2022). Readability scores used Dale-Chall Readability Score (DCRS) and Flesch-Kincaid Grade Level (FKGL) measures.

Table 1 compares *LED*, a baseline model that is only trained on the original documents to generate

	DCRS	FKGL	bartscore	summac	bertscore	rouge1	rouge2	rougeL
LED	12.36	15.51	**-2.21**	**21.37**	86.22	45.01	15.44	24.15
All	12.29	**15.42**	-2.28	21.36	**86.26**	**45.31**	**15.75**	**24.47**
Wikipedia	**12.27**	15.58	-2.24	21.17	86.15	44.70	15.01	23.81
Abstracts	12.30	15.62	-2.29	21.30	86.18	44.67	15.29	24.04
UMLS	12.32	15.61	-2.23	21.22	86.14	44.67	14.94	23.80
Wiki Simple	12.31	15.58	-2.22	21.24	86.14	44.51	14.87	23.74

Table 1: Readability, factuality, and relevancy scores for different grounding sources compared against LED.

the output lay summaries, and *All*, where the model is trained on the original document and supplemented with passages from the retrieval sources.[3]

The results in table 1 show a few trends, first that *All* improves relevancy and readability over the *LED* setting. However, these gains are not very large which indicates that the grounding sources were not a very important signal for the model. Additionally, despite our hypothesis, factuality is not improved with grounding.

3.2 Analysis of Grounding Methods

To assess the impact of various retrieval sources on summarization quality, we trained a model for each retrieval corpus (table 1). The most noticeable difference is that articles grounding in passages from *Scientific Abstracts* and *Wikipedia* have the highest relevancy scores. Possibly due to a similarity in the text distributions of these sources with reference lay summaries. Other grounding sources have different genre roles such as definitions in the case of *UMLS* or non-expert reference literature in the case of *Wikipedia Simple*. All methods generally improve readability as measured by DCRS with *Wikipedia* having the largest effect.

[3]Our method *All* was submitted to the BioLaySumm Task 1 see appendix A for the results.

3.3 Zero-shot Summarization

The lack of differences between grounding sources inspired us to consider an experiment where a model might be more likely to use grounding sources. We designed a zero-shot summarization experiment (with GPT 3.5) using the same method from section 2 but with 2,048 tokens selected for the original document and the rest from the original article (see Appendix D for more details).

The results in table 2 show that grounding sources make much more of a difference for zero-shot summaries than trained summaries. The *Without* input, which means without any grounding sources, has the highest relevancy and factuality scores indicating that in a zero-shot setting grounding sources tend to provide some distraction. Interestingly, *UMLS* and *Wikipedia Simple* encourage more readable summaries than other methods which is what we would expect to find since *UMLS* provides definitions which are vital to assisting non-experts with engaging scientific findings and *Wikipedia Simple* provides plain language encyclopedic knowledge designed to be readable. As we saw above, *Scientific Abstracts* as a grounding source allows us to construct more relevant summaries, perhaps due to *Scientific Abstracts* preserving the scientific language and context that is still vital for lay summaries.

	DCRS	FKGL	bartscore	summac	bertscore	rouge1	rouge2	rougeL
All	11.31	14.04	-3.35	**20.30**	85.78	39.09	11.26	21.50
Without	11.23	14.15	**-3.08**	**20.30**	**86.02**	**40.66**	**11.98**	**21.97**
Abstracts	11.45	14.48	-3.39	20.26	85.72	39.41	11.42	21.52
UMLS	**10.80**	**13.42**	-3.37	20.25	85.56	38.40	10.00	20.69
Wikipedia	11.35	14.11	-3.49	20.26	85.49	39.05	10.48	20.61
Wik Simple	10.88	13.43	-3.75	20.21	84.96	36.40	9.03	19.20

Table 2: Zero-shot summarization setting exploring summary quality for each grounding source

4 Discussion

Grounding the original document in retrieval results helped most with readability, marginally with relevance and not at all with factuality. We believe this gives credence to the idea that grounding for lay summarization is primarily helpful for providing the model with background information that helps explain concepts, define terms, and otherwise situate the reader with the necessary information to be able to understand a scientific finding. We saw that the *UMLS* and *Wikipedia Simple* sources provided best effect on improving readability in the zero-shot summarization setting which is intuitive in that they provide clear definitions and encyclopedic background knowledge in simple terms. *Scientific Abstracts* had the best effect of all sources on relevancy which is also intuitive because related abstracts such as ones with similar findings might help construct a more robust summary. These results indicate that we should continue to investigate the role of background information on improving readability and relevance in lay summarization. In particular, our retrieval method is quite simple and more sophisticated methods of retrieval such as dense passage retrieval (Izacard et al., 2022) could be used to enhance the relevancy of grounding documents. Additionally, future work should investigate methods that learn more strongly to take adavantage of grounding sources.

The lack of improvement in factuality is an important aspect that future work should investigate. One explanation is that grounding sources could introduce factual or relevancy errors if the retrieved documents are irrelevant or incorrect and they end up being used in the generated summary. However, factuality metrics only measure the summary against original document, this means that statements that cannot be grounded in the original document may be penalized. This is an issue in lay summarization where there can be statements that are factual and provide necessary background knowledge but cannot be found in the original document. Future work should investigate methods of measuring factuality that are able to incorporate the necessary background knowledge when measuring the factuality of a lay summary.

5 Related Work

There are a number of works looking at automatic lay summarization in the biomedical domain Goldsack et al. (2022); Guo et al. (2022); Luo et al.

(2022); Devaraj et al. (2021). One central issue is understanding the effect of text simplification on various aspects of summary quality such as relevancy. (Devaraj et al., 2022) explores the effect of text simplification on factuality by introducing a taxonomy of different error types that allows them to observe that while factual errors of missing information are a common error across generated and gold summaries, errors of substitution such as mixing up entities is a common occurrence in text simplification summarization models.

Supplementing source documents with external knowledge has been one of the main interventions discussed for mitigating factual errors and hallucinations in natural language generation systems (Shuster et al., 2021; Lewis et al., 2021). With the idea that having access to grounding sources allows models to draw on those sources when generating text rather than being forced to rely on parametric knowledge which may be flawed (Thoppilan et al., 2022; Mallen et al., 2022). In summarization, researchers have used factual knowledge from external sources to improve factuality by encoding external knowledge in models during training (Zhu et al., 2021; Mao et al., 2022) or to correct already generated summaries (Lee et al., 2022).

Guo et al. (2022) evaluates retrieval augmentation as a method for enhancing abstracts with background knowledge. They use definitions from UMLS and Wikipedia as different retrieval corpora and find their method improves both the readability and relevancy of summaries while maintaining a similar level of factuality as models not using grounding. However, they did not perform retrieval-augmented generation for the lay summary generation task which is the novel contribution in this work.

Limitations

Retrieval augmentation adds complexity to natural language generation requiring a separate retrieval module before the text generation step can begin. Additionally, retrieval augmentation possibly introduces more input text than the original input which is problematic for many neural network architectures with limited input space especially in the case of summarizing entire scientific papers. Finally, retrieval augmentation itself could introduce factual or relevancy errors if the retrieved documents are irrelevant or incorrect and they end up being used in the generated summary.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. ArXiv:2004.05150 [cs].

Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. 2021. Paragraph-level Simplification of Medical Texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4972–4984, Online. Association for Computational Linguistics.

Ashwin Devaraj, William Sheffield, Byron C. Wallace, and Junyi Jessy Li. 2022. Evaluating Factuality in Text Simplification.

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2021. GO FIGURE: A Meta Evaluation of Factuality in Summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 478–487, Online. Association for Computational Linguistics.

Tomas Goldsack, Zheheng Luo, Qianqian Xie, Carolina Scarton, Matthew Shardlow, Sophia Ananiadou, and Chenghua Lin. 2023. Overview of the biolaysumm 2023 shared task on lay summarization of biomedical research articles. In *Proceedings of the 22st Workshop on Biomedical Language Processing*, Toronto, Canada. Association for Computational Linguistics.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yue Guo, Wei Qiu, Gondy Leroy, Sheng Wang, and Trevor Cohen. 2022. CELLS: A Parallel Corpus for Biomedical Lay Language Generation. ArXiv:2211.03818 [cs].

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177. Place: Cambridge, MA Publisher: MIT Press.

Hwanhee Lee, Cheoneum Park, Seunghyun Yoon, Trung Bui, Franck Dernoncourt, Juae Kim, and Kyomin Jung. 2022. Factual error correction for abstractive summaries using entity retrieval. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 439–444, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. ArXiv:2005.11401 [cs].

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. Readability controllable biomedical document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4667–4680, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. When not to trust language models: Investigating effectiveness and limitations of parametric and non-parametric memories.

Qianren Mao, Jianxin Li, Hao Peng, Shizhu He, Lihong Wang, Philip S. Yu, and Zheng Wang. 2022. Fact-Driven Abstractive Summarization by Utilizing Multi-Granular Multi-Relational Knowledge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1665–1678. Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval Augmentation Reduces Hallucination in Conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana,

Dominican Republic. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating Generated Text as Text Generation. ArXiv:2106.11520 [cs].

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2021. Enhancing Factual Consistency of Abstractive Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 718–733, Online. Association for Computational Linguistics.

A BioLaySumm

The BioLaySumm (Goldsack et al., 2022) dataset has two splits, one for articles and lay summaries from PLOS and the other for articles and summaries from eLife. There are 29,119 training samples (24,773 from PLOS and 4,346 from eLife) and 1,617 validation samples (1,376 from PLOS and 241 from eLife). There is an additional test set but this was not used as the reference summaries are not publically available. All results are presented using the validation set. Our zero-shot summarization experiment uses a random sample of 300 summaries from the PLOS validation sample set.

Our method *All* was submitted to the first shared task of BioLaySumm (Goldsack et al., 2023) resulting in 13th place overall out of 21 submissions. See table 3.

B Model Training and Inference

In our experiments, we use the Longformer Encoder-Decoder (LED) (Beltagy et al., 2020). We use the `allenai/led-base-16384` checkpoint on huggingface. All models are trained using 16,384 input tokens and an output length of 512 tokens except for *LED 8K* which is trained using 8,192 input tokens. All models are trained on 4 v100 GPUs. Models are trained for 4 epochs and evaluated every 5,000 steps. The checkpoint with the best Rouge2 is selected as our best model. We set global attention to the first token and on a `<|SEARCH|>` token the model if the source document was supplemented with retrieval results. For inference, we sample text from the model using the following parameters with greedy decoding: number of beams 4, min length 100, length penalty 2.0, early stopping True, no repeat ngram size 3. We present LED 8k as the baseline comparison model in the main paper so that all models use the same amount of original article tokens. For a picture of what LED trained with 16k tokens from the original article see LED 16k in table 4.

The crossencoder we used for re-ranking passages according to semantic similarity with the first 1,024 tokens of the original document was `cross-encoder/ms-marco-MiniLM-L-6-v2` using the sentence transformers library (Reimers and Gurevych, 2019).

C Grounding Sources

APA reference strings were constructed using Crossref[4] using the first author name, title, and year from Crossref metadata matched on the BioLaySumm document IDs. The UMLS definitions were from the sciscapcy entity linker (Neumann et al., 2019). The Wikipedia index was a prebuilt index `enwiki-paragraphs` from Pyserini. We constructed the Wikipedia Simple index using the march 1st dump [5]. The index was constructed extracting the plain text from each article and chunking articles into 6 sentence chunks. The 6 sentence chunks were indexed with Pyserini (Lin et al., 2021). The Scientific Abstracts index was an Elasticsearch index of Crossref abstracts, it was accessed through scite.ai search api [6]. Table 5 illustrates the distribution of the number of passages that have been selected for the grounding context of the *Grounded* model in table 1.

---

[4]https://www.crossref.org
[5]https://dumps.wikimedia.org/simplewiki/20230301/
[6]https://api.scite.ai/docs

FKGL	DCRS	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	BARTScore
12.73 (12)	10.50(18)	47.69 (11)	16.76 (1)	44.30 (11)	86.01 (5)	-2.32 (13)

Table 3: Results of out model on the test set for BioLaySumm Shared Task 1. The number in parenthesis indicates how our model did compared to other submissions. Our model did best on Rouge 2 scores and placed 13th overall out of 21 submissions.

	DCRS	FKGL	bartscore	summac	bertscore	rouge1	rouge2	rougeL
LED 8k	12.36	15.51	**-2.21**	**21.37**	86.22	45.01	15.44	24.15
LED 16k	**12.24**	**15.32**	-2.22	21.34	**86.26**	**45.35**	**15.58**	**24.35**
All	12.29	15.42	-2.28	21.36	**86.26**	45.31	15.75	24.47

Table 4: Readability, factuality, and relevancy scores for our methods measured against a LED baseline trained with 8k and 16k tokens of the paper to be summarized respectively. Lower is better for DCRS and FKGL and higher is better for the other scores.

Source	Mean	STD
Abstracts	8.81	3.0
Wikipedia	6.66	2.4
UMLS	2.56	2.1
Wiki Simple	0.71	0.8

Table 5: Distribution of grounding sources selected by our method (section 2). Scientific Abstracts and Wikipedia are selected far more often than other sources.

D Zeroshot summarization

For our zero-shot summarization setting we use `GPT-3.5-turbo` with a temperature of 0. We evaluate on the entire BioLaySumm so we can compare with results in table 1. We prompt the model with the following prompts:

System Prompt:
You are a document summarizing agent focusing on summarizing documents to make them readable for a lay audience. Summarize the documents presented by the user in as simple terms as possible.
Prompt:
Summarize this document for a lay audience:
{document}
Below are a set of search results that ground the above document.
{search results}

E Sample Summaries

The below are samples of generated summaries of the article with the ID journal.pgen.1002882 from the PLOS validation set. *Original* indicates the original summary.

Original:

Messenger RNAs carry the instructions necessary to synthesize proteins that do work for the cell . Extending beyond the protein-coding sequence of a given mRNA is an additional stretch of sequence , harboring signals that govern how much protein is made and how long the mRNA remains in the cell before it is broken down . The incorporation of this end region into mature mRNA is itself subject to change; for the vast majority of human genes , how and why cells use different mRNA ends remains largely unknown . In this work , we surveyed mRNA ends from 10 , 000 genes in immune cells from genetically distinct human individuals . We found that mRNA end positions were not randomly distributed , but rather preferentially flanked the locations of regulatory signals that govern mRNA fate . The usage of these mRNA length forms and regulatory elements varied across individuals and could be dissected molecularly . Our results uncover key mechanisms and regulatory effects of transcript end processing , particularly as these are perturbed by genetic differences between humans .

Grounded:

Gene expression varies widely between individuals of a population, and regulatory change can underlie phenotypes of evolutionary and biomedical relevance. A key question in the field is how DNA sequence variants impact gene expression, with most mechanistic studies to date focused on the effects of genetic change on regulatory regions upstream of protein-coding sequence. By contrast, the role of RNA 3-end processing in regulatory variation remains largely unknown, owing in part to the challenge of identifying functional elements in 3 untranslated regions. In this work, we conducted a genomic survey of transcript ends in lymphoblastoid cells derived from genetically distinct human individuals. Our analysis mapped the cis-regulatory architecture of 3 gene ends, finding that transcript end positions did not fall randomly in untranslation regions, but rather preferentially flanked the locations of regulatory elements, including miRNA sites. The usage of these transcript length forms and motifs varied across human individuals, and polymorphisms in polyadenylation signals and other 3 motifs were significant predictors of expression levels of the genes in which they lay. Our findings suggest that the strategy of integrating observed 3 transcript ends with inferred 3 regulatory motifs will prove to be a critical tool in continued efforts to interpret human genome variation.

LED:

Gene expression varies widely between individuals of a population, and regulatory change can underlie phenotypes of evolutionary and biomedical interest. A key question in the field is how DNA sequence variants impact gene expression, with most mechanistic studies to date focused on the effects of genetic change on regulatory regions upstream of protein-coding sequence. By contrast, the role of RNA 3-end processing in regulatory variation remains largely unknown, owing in part to the challenge of identifying functional elements in 3 untranslated regions. In this work, we conducted a genomic survey of transcript ends in lymphoblastoid cells from genetically distinct human individuals. Our analysis mapped the cis-regulatory architecture of 3 gene ends, finding that transcript end positions did not fall randomly in untranstranslated regions, but rather preferentially flanked the locations of regulatory elements, including miRNA sites. The usage of these transcript length forms and motifs varied across human individuals, and polymorphisms in polyadenylation signals and other 3 motifs were significant predictors of expression levels of the genes in which they lay. Our results underscore the importance of 2 mechanisms at play in the genetics of gene expression: the usage of distinct 3 end processing signals and the effects on 3 sequence elements that determine transcript fate. Our findings suggest that the strategy of integrating observed transcript ends with inferred 3 regulatory motifs will prove to be a key component in the systems-level modeling of regulatory networks and their variation.

Zero-shot UMLS:

This document talks about how genetic differences between individuals can affect how genes are expressed, which can have important implications for evolution and medicine. Specifically, the document focuses on how differences in the way RNA molecules are processed at their ends can impact gene expression. The researchers used a technique called 3'-end RNA-seq to study this process in human cells, and found that variations in the processing of RNA ends can be influenced by genetic differences between individuals. The study sheds light on the regulatory elements that control gene expression and suggests that understanding these elements will be important for interpreting human genome variation.