LMU at PerAnsSumm 2025: LlaMA-in-the-loop at Perspective-Aware Healthcare Answer Summarization Task Factuality

Tanalp Agustoslu
LMU Munich
t.agustoslu@campus.lmu.de

Abstract

In this paper, we describe our submission for the shared task on Perspective-aware Healthcare Answer Summarization. Our system consists of two quantized models of the LlaMA family, applied across fine-tuning and few-shot settings. Additionally, we adopt the SumCoT prompting technique to improve the factual correctness of the generated summaries. We show that SumCoT yields more factually accurate summaries, even though this improvement comes at the expense of lower performance on lexical overlap and semantic similarity metrics such as ROUGE and BERTScore. Our work highlights an important trade-off when evaluating summarization models.

1 Introduction

In this paper, we present our submission for the shared task on Perspective-aware Healthcare Answer Summarization (PerAnsSumm) (Agarwal et al., 2025). PerAnsSumm comprises two tasks: span identification and summarization. Given a medical question-answer pair as input, the system must identify spans within the answer and classify them into five distinct perspectives: 'cause,' 'suggestion,' 'experience,' 'question,' and 'information.' In Task 2, the system utilizes these extracted perspective categories to generate summaries corresponding to the same five perspectives. The final summaries encompass all perspectives present in the given answer within the QA pair.

The shared task leverages the PUMA dataset (Naik et al., 2024), a perspective-aware annotated corpus of QA pairs and their respective summaries extracted from Yahoo!'s L6 corpus. Participants are provided with annotated spans and summaries in the training and development sets, while the test set contains only QA pairs. The first task, span identification, is evaluated at the lexical level using strict and proportional matching metrics[1]. The second task, summarization, is assessed using relevance metrics, ROUGE (Lin, 2004), BERTScore (Zhang et al., 2019), METEOR (Banerjee and Lavie, 2005) and BLEU (Papineni et al., 2002) at both lexical and semantic levels. Additionally, the organizers introduce two metrics, AlignScore (Zha et al., 2023) and SummaC (Laban et al., 2022), to evaluate the factuality of generated summaries. We participate in Task 2.2 (Factuality) of the shared task, where we approach the problem by leveraging two quantized models from the LLaMA family (Grattafiori et al., 2024) in fine-tuning, few-shot and chain-of-thought (CoT) (Wei et al., 2023) prompting settings. Depending on the approach, we either generate summaries directly or first identify spans and then incorporate them into the summarization process.

2 Related Work

The prominence of Large Language Models (LLMs) in the medical domain has been well documented through surveys and evaluation benchmarks in recent years. Integrating them with various prompting strategies, such as zero-shot, few-shot, CoT, and Analogical Reasoning (Yasunaga et al., 2024), has yielded promising results (Vatsal and Singh, 2024; Liévin et al., 2023; Jullien et al., 2023). Their ability to handle long contexts in medical domain and leverage intermediate reasoning steps make them suitable candidates not only for text summarization but also for information extraction tasks such as named entity recognition or event extraction (Xu et al., 2024; Bian et al., 2023; Yuan et al., 2023).

The effectiveness of LLMs in these tasks, however, is closely tied to their scale. Kaplan et al. (2020) introduced the concept of sample efficiency as part of their scaling laws, showing that larger

[1] https://github.com/PerAnsSumm/Evaluation/blob/main/eval.py

Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), pages 380–388
May 4, 2025

neural language models require fewer optimization steps and are more sample efficient than their smaller counterparts. This suggests that, even with a small to moderate-sized datasets, opting for a larger model can be advantageous. However, a key limitation of LLMs is their computational cost, which restricts their deployment in resource-constrained environments. To address this, low-rank adaptation (LoRA) method has been proposed (Hu et al., 2021). LoRA freezes the pre-trained model weights and updates only low-rank approximations of the weight matrices. This drastically reduces the number of trainable parameters, thereby significantly lowering computational overhead. QLoRA (Dettmers et al., 2023) further optimizes this approach by quantizing the model weights typically to 4-bit precision while utilizing paged optimizers to efficiently manage memory, avoiding spikes by dynamically offloading data between GPU and CPU memory.

In our work, we employ quantized versions of LLaMA-70B and LLaMA-8B from the Unsloth library[2] and explore few-shot as well as fine-tuning settings. Additionally, we incorporate a variation of CoT prompting called Summary Chain-of-Thought (SumCoT) (Wang et al., 2023), which is inspired by Lasswell's Communication Model (Laswell, 1948) and designed for element extraction and text summarization tasks in an end-to-end manner.

Dataset Statistics	**Dev Set**	**Train Set**
Total Instances	959	2236
Total Tokens	239,486	555,249
Avg Tokens per Instance	249.72	248.32
Avg Words per Instance	216.02	214.78
Avg Answers per Instance	3.23	3.11
Avg Perspectives per Instance (Answers)	1.97	1.97
Avg Perspectives per Instance (Summaries)	1.96	1.95
Perspective Distribution (Answers)		
EXPERIENCE	316	747
INFORMATION	735	1767
CAUSE	139	308
SUGGESTION	595	1360
QUESTION	102	215
Perspective Distribution (Summaries)		
EXPERIENCE	315	745
INFORMATION	733	1742
CAUSE	138	305
SUGGESTION	595	1363
QUESTION	101	213

Table 1: PUMA Dataset Statistics for Development and Training Sets. Test Set consists of 50 instances and only includes QA pairs with a context information without providing any perspective spans or summaries.

3 Methods

We evaluate a set of prompting strategies to generate factually correct summaries. Our approaches include fine-tuning, few-shot, and SumCoT prompting. As a baseline, we use LLaMA-8B with fine-tuning.

3.1 Fine-Tuning

For fine-tuning, we use the training dataset provided by the organizers and employ the 4-bit quantized LLaMA-8B model with a learning rate of 2e-4 and train it for 3.5 epochs. Additionally, we configure all applicable modules with a rank of 16 and an alpha value of 16.

3.2 Few-Shot

For few-shot prompting, we use a quantized LLaMA-8B model in a 1-shot setting, where in-context examples are randomly selected for each inference to prevent the model from overfitting to a fixed set of examples. Each example includes both labeled spans and their corresponding summaries, and the model is instructed to generate only the summary. The model used in this setting has already been fine-tuned on the provided training set.

3.3 Summary Chain-of-Thought (SumCoT)

We incorporate a variant of CoT prompting called SumCoT, which is designed for element extraction and text summarization tasks in an end-to-end manner. This approach is inspired by Lasswell's Communication Model, which later found itself application in journalism as the 5W framework (Who, What, When, Where, Why). Following prior work by Wang et al. (2023) that suggests that performance gains become evident only at scale, we employ a 4-bit quantized version of LLaMA-70B. In line with their findings, we formulate our questions using only a single type of W-question, specifically "What", as it can encapsulate the essence of all other questions.[3] We later append the five distinct perspectives found in our dataset to the questions. As we observe the stabi-

[2] https://huggingface.co/unsloth

[3] https://github.com/Alsace08/SumCoT/blob/master/prompts/cot_element_extraction.txt

Prompt Template
You are provided with a text containing community-based questions and answers from the medical domain. Your task is to analyze the answers by identifying and considering different perspectives such as 'Information', 'Cause', 'Suggestion', 'Experience', and 'Question' as in the provided examples below and then summarize the text into a coherent summary. Only output the summaries and nothing else.
In-Context Examples:
Example: **Question:** Do I have lupus? **Context:** I had a fever and fatigue. I looked at the symptoms on the internet. My doctor disagrees with me. [...] **Answers:** What other symptoms did you have? It's usually never lupus. Listen to your doctor. Lupus is an autoimmune disorder [...] **EXPERIENCE_GROUP:** My teacher used to say this. It turns out it was just a flu. **INFORMATION_GROUP:** Lupus is an autoimmune disorder [...] **CAUSE_GROUP:** I had a fever and fatigue. [...] **SUGGESTION_GROUP:** Listen to your doctor. [...] **QUESTION_GROUP:** What other symptoms did you have? [...] **EXPERIENCE_SUMMARY:** In users experience, [...] **INFORMATION_SUMMARY:** For information purposes [...] **CAUSE_SUMMARY:** Some of the causes are [...] **SUGGESTION_SUMMARY:** It's suggested that [...] **QUESTION_SUMMMARY:** It's inquired [...]
Text: {text}
Answer: {answer}

Table 2: Prompt Template for Few-Shot Method. Summary examples are given with common start phrases found in the PUMA dataset.

lizing effect of it during generations, we additionally prefix the phrase "Let's think step by step." (Kojima et al., 2023) before the model extracts the relevant perspectives. After eliciting information about spans from the model, we then provide the fine-tuned 8B model with the output generations of the 70B variant and let it generate summaries based on the extracted perspectives.

Prompt Template
You are provided with a text containing community-based questions and answers from the medical domain. Your task is to analyze the answers by identifying and considering different perspectives such as 'Information', 'Cause', 'Suggestion', 'Experience', and 'Question'. Show your reasoning steps while extracting.
Questions: What are the important suggestions in these answers? What are the important causes in these answers? What are the important informations in these answers? What are the important questions in these answers? What are the important experiences in these answers? Please answer the above questions.
Text: {text}
Answer: Let's think step by step. {answer}

Table 3: Prompt Template for SumCoT Method

4 Evaluation Protocol

The PerAnsSumm shared task evaluates submissions across three axes. Task 1 focuses on lexical overlap, using both proportional and strict matching metrics to assess the accuracy of extracted label spans from answers as well as the generated summaries. Task 2 is further divided into two subcategories: Task 2.1 evaluates lexical and semantic similarity using relevance metrics, ROUGE, BERTScore, METEOR and BLEU. Task 2.2 assesses the factual consistency of the generated spans and summaries using AlignScore and SummaC.

AlignScore is a reference based metric, formally:

$$\text{AlignScore}(x, y) = \frac{1}{|x|} \sum_{i=1}^{|x|} \max_j s(x_i, y_j) \quad (1)$$

where x is the generation, y is the reference and $|x|$ is the number of sentences in the generation, and $\max_j s(x_i, y_j)$ selects the maximum alignment score for each sentence of the generation across all chunks of the reference (split into approximately 350-token chunks for RoBERTa (Liu et al., 2019)) using an unified alignment function trained on a diverse set of NLP tasks (e.g., natural language inference, question answering, semantic similarity, fact verification) with a combined dataset of 4.7 million examples.

SummaC follows a similar chunking approach, but adds an additional layer by using an NLI model to scan sentence pairs. These entailment scores are aggregated into histogram bins, which are then processed through a convolutional neural network (CNN) (LeCun and Bengio, 1998) to produce scalar values for each summary sentence. These scalar values are averaged to compute the final consistency score.

Despite the significant drawbacks of frequent test set evaluation (van der Goot, 2021), we evaluated our approaches on the test set due to time constraints, as the hyperparameters for AlignScore and SummaC were not known until a later stage of the shared task.

5 Results

The results presented in Table 4 provide insights into the impact of different methods on improving factuality and help address our research question: *Can we improve the factuality of generated summaries with in-context-learning and chain-of-thought prompting?*

Table 4 shows that there is no clear winner across all metrics. The standard fine-tuning method achieves the best results in relevance metrics, with the exception of the few-shot approach,

Name	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	METEOR	BLEU	Rel. Avg.	AlignScore	SummaC	Fact. Avg.
Fine-Tuning	0.2550	0.0991	0.2288	**0.6448**	0.2349	0.0643	**0.2545**	0.3235	0.2398	0.2817
Few-Shot	0.1912	0.0573	0.1701	**0.6512**	0.1636	0.0489	0.2137	0.2263	0.2262	0.2263
8B-Labels	0.2226	0.0896	0.2044	0.5413	0.2045	0.0704	0.2221	0.3246	0.2274	0.2760
70B-Labels (SumCoT)	0.2148	0.0905	0.1942	0.5351	0.2032	0.0595	0.2162	**0.3564**	**0.2471**	**0.3017**

Table 4: PerAnsSumm 2025 test set results for all evaluated approaches. All approaches use the same fine-tuned model for summary generation. *Few-Shot* used in a 1-shot setting. In the *8B-Labels*, spans are identified by the fine-tuned 8B model and the output passed to the same 8B fine-tuned model for summary generation. In the *70B-Labels (SumCoT)*, spans are identified by 70B model without fine-tuning and the output passed to the same 8B fine-tuned model for summary generation. ROUGE scores measure n-gram overlap, BERTScore evaluates semantic similarity, METEOR compares unigrams, synonyms and stemming with penalties for word order differences, BLEU compares n-gram precision between the generated summary and the ground truth, applying a brevity penalty for shorter generations. AlignScore and SummaC measure factual consistency. *Rel. Avg* shows the average of ROUGE, BERTScore, METEOR and BLEU, and *Fact. Avg.* shows the average of AlignScore and SummaC.

which surpasses fine-tuning in semantic similarity when evaluated using contextual BERT (Devlin et al., 2019) embeddings. However, the few-shot approach exhibits relatively low ROUGE scores (especially ROUGE-2) alongside lower METEOR and BLEU scores. This results in a higher average relevance score for fine-tuning, suggesting that the model may have prioritized the in-context examples while being penalized for differences in word order and shorter generations by METEOR and BLEU during few-shot generations. A similar pattern is observed in ROUGE-L, where the longest common subsequence between the generated and reference summaries is less aligned. When it comes to factuality, surprisingly, the few-shot approach does not lead to any improvements and performs significantly worse than the standard fine-tuning method. Additionally, we observe a slight decline in SummaC and average factuality with the 8B label extraction method, along with a notable drop in BERTScore. It appears that in both approaches, the model was biased toward the in-context examples and the extracted spans, respectively. Moreover, the extracted spans from the fine-tuned model may be incorrect, as the model was trained solely for the summary generation task. This suggests that it may be heavily relying on its memorized knowledge of training set labels acquired during parameter updates, which could have skewed the metrics.

On the other hand, even without any fine-tuning, the SumCoT approach with the 70B label extraction method shows a noticeable impact. Despite a significant drop in BERTScore and ROUGE (similar to the 8B label extraction) the final summaries are the most factually accurate. This also highlights the important trade-off between relevance and factuality metrics when evaluating summarization models. Lexical and semantic alignment does not always guarantee hallucination-free, factually correct summaries.

The challenge of identifying the optimal summary is a complex and nuanced issue. As proven by Schluter (2017), performing a ROUGE evaluation of a summarization model for optimal summaries is an NP-hard task and relying solely on relevance metrics does not capture the full capabilities of the implemented system. As demonstrated in this shared task, it makes sense to introduce multiple perspectives into the evaluation by incorporating additional metrics and averaging them to mitigate the shortcomings of any single metric.

6 Conclusion

In our submission, we explored several approaches to improve the factuality of generated summaries. Our best-performing method, SumCoT, involved extracting spans using a 4-bit quantized LLaMA 70B model with W-Questions, and feeding the output into a fine-tuned 8B model to generate summaries. This approach led to improvements in the factuality of the generated summaries compared to standard fine-tuning and few-shot methods. However, these improvements are not always reflected in relevance metrics such as ROUGE and BERTScore. Our final submission ranks 15th in AlignScore, 16th in SummaC, and 15th in average factuality on the official leaderboard[4].

[4] https://docs.google.com/spreadsheets/d/1faysHdA7YQ-xELztsm7jA5RPTMh7lP7tycsjd8ANLGE/

7 Limitations

In this section, we highlight some shortcomings of our implemented system and outline potential directions for future work.

One notable limitation in our approach is the choice of random sampling for the few-shot examples, which was intended to prevent bias toward the same examples. However, Gema et al. (2024) demonstrates the effectiveness of the BM25 retriever over naive random sampling. BM25 allows for the selection of only the most relevant in-context examples, which could improve performance in future iterations of the shared task.

Another limitation is our use of quantization due to computational constraints, which may have affected our findings. As highlighted by Pochinkov (2024), performance degradation is often inevitable in quantized LLaMA models.

Our final submission, SumCoT, showed improvements in factuality metrics. However, as noted by Wang et al. (2023), the success of the proposed approach is often correlated with the model's parameter count. We expect that using larger models, including closed-source ones like GPT (OpenAI et al., 2024), would likely amplify these results. An important consideration, however, when transitioning to closed-source models, is the memorization ability of neural language models (Carlini et al., 2023) and the issue of data leakage. Balloccu et al. (2024) identified potentially leaked datasets within the training data of ChatGPT and GPT-4 by systematically reviewing 255 research papers. In our case, as the PUMA dataset and Yahoo's L6 Corpus are not publicly available and primarily cover texts from the early 2000s to early 2010s, data leakage is unlikely to be a significant concern. However, taking basic measures and implementing simple n-gram matching metrics to detect potential data leakage in model completions of any given data instance (Gema et al., 2024) or adopting the Contamination Detection via Output Distribution (CDD) framework proposed by Dong et al. (2024) could further strengthen the reliability of the obtained results and would align well with the broader goal of trustworthy AI.

References

Siddhant Agarwal, Md Shad Akhtar, and Shweta Yadav. 2025. Overview of the peranssumm 2025 shared task on perspective-aware healthcare answer summarization. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health) @ NAACL 2025*, Albuquerque, USA. Association for Computational Linguistics.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Junyi Bian, Jiaxuan Zheng, Yuyi Zhang, and Shanfeng Zhu. 2023. Inspire the large language model by external knowledge on biomedical named entity recognition.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *CoRR*, abs/2305.14314.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12039–12050, Bangkok, Thailand. Association for Computational Linguistics.

Aryo Gema, Giwon Hong, Pasquale Minervini, Luke Daines, and Beatrice Alex. 2024. Edinburgh clinical NLP at SemEval-2024 task 2: Fine-tune your model unless you have access to GPT-4. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1894–1904, Mexico City, Mexico. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, 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, Danny Wyatt, 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, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, 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, Jack Zhang, 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, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, 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, Maria Tsimpoukelli, 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, Ning Zhang, 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, Rohan Maheswari, 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, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, 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, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, 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, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, 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, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, 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, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, 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, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao

Liu, 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, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, 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, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, 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, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, 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, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Maël Jullien, Marco Valentino, Hannah Frost, Paul O'regan, Donal Landers, and André Freitas. 2023. SemEval-2023 task 7: Multi-evidence natural language inference for clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2216–2226, Toronto, Canada. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

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.

Harold D Laswell. 1948. The structure and function of communication in society.

Yann LeCun and Yoshua Bengio. 1998. *Convolutional networks for images, speech, and time series*, page 255–258. MIT Press, Cambridge, MA, USA.

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

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

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2023. Can large language models reason about medical questions?

Gauri Naik, Sharad Chandakacherla, Shweta Yadav, and Md Shad Akhtar. 2024. No perspective, no perception!! perspective-aware healthcare answer summarization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15919–15932, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Nicky Pochinkov. 2024. Comparing quantized performance in llama models. Last access: 02.25.2025.

Natalie Schluter. 2017. The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45, Valencia, Spain. Association for Computational Linguistics.

Rob van der Goot. 2021. We need to talk about train-dev-test splits. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4485–4494, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shubham Vatsal and Ayush Singh. 2024. Can GPT redefine medical understanding? evaluating GPT on biomedical machine reading comprehension. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 256–265, Bangkok, Thailand. Association for Computational Linguistics.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024. Large language models for generative information extraction: A survey.

Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H. Chi, and Denny Zhou. 2024. Large language models as analogical reasoners.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. Zero-shot temporal relation extraction with ChatGPT. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 92–102, Toronto, Canada. Association for Computational Linguistics.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert.