AICOE at PerAnsSumm 2025: An Ensemble of Large Language Models for Perspective-Aware Healthcare Answer Summarization

Rakshith R, Mohammad Sameer Khan, Ankush Chopra
AICOE, Tredence
Bengaluru, India
{rakshith.r, mohammed.sameerkhan, ankush.chopra}@tredence.com

Abstract

The PerAnsSumm 2024 shared task at the CL4Health workshop focuses on generating structured, perspective-specific summaries to enhance the accessibility of health-related information. Given a Healthcare community QA dataset containing a question, context, and multiple user-answers, the task involves identifying relevant perspective categories, extracting spans from these perspectives, and generating concise summaries for the extracted spans. We fine-tuned open-source models such as Llama-3.2 3B, Llama-3.1 8B, and Gemma-2 9B, while also experimenting with proprietary models including GPT-4o, o1, Gemini-1.5 Pro, and Gemini-2 Flash Experimental using few-shot prompting. Our best-performing approach leveraged an ensemble strategy, combining span outputs from o1 (CoT) and Gemini-2 Flash Experimental. For overlapping perspectives, we prioritized Gemini. The final spans were summarized using Gemini, preserving the higher classification accuracy of o1 while leveraging Gemini's superior span extraction and summarization capabilities. This hybrid method secured fourth place on the final leaderboard among 100 participants and 206 submissions.

1 Introduction

In recent years the widespread adoption of social media has sprung up various community question answer forums especially in the medical domain. Users often rely on others experience or suggestions. They post a query along with information as context and multiple users can answer them. The answers vary in multiple aspects depending on the user's question, the experience of the person replying etc. Hence traditional summarization techniques are not particularly useful since they combine everything. User's answers include multiple perspectives and the aim of this shared task (Agarwal et al., 2025) is to identify them and form more meaningful summaries for users to make more informed healthcare decisions. The perspectives are 'Cause', 'Suggestion', 'Experience', 'Question', and 'Information'. An example is displayed in Figure 1. The recent rise of Large Language Models enable much more accurate perspective identification and summarization than traditional transformers. We leverage these LLM's both proprietary and open source for the task. We finetune open-source smaller models like Llama 3b, 8b (Grattafiori et al., 2024) and Gemma 9b (Team et al., 2024) for the task. We observe that finetuning significantly improves the base models performance on the task and even outperforms models like GPT 4o (8 shot prompt) (OpenAI et al., 2024).

2 Related Work

Span prediction and Abstractive Summarization are popular tasks in the ML domain for a long time. Transformer models have been used ever since the Transformer paper (Vaswani et al., 2023). Models like BERT (Devlin et al., 2019), Roberta (Liu et al., 2019) and it's variants were the best performing models of their time. This was soon followed by pre-trained language models (PLMs) like BART (Lewis et al., 2019), T5 (Raffel et al., 2023), PEGASUS (Zhang et al., 2020) etc.which achieved state of the art results in their time.

In the medical domain these models were trained on biomedical corpora like PubMed and MIMIC-III giving to rise of domain specific pre-trained language models (PLMs) like BioBERT (Lee et al., 2019), BioBART (Yuan et al., 2022), and clinicalBERT (Huang et al., 2020) which did much better in medical domain tasks. There are efforts in summarizing diverse types of content, including biomedical literature using these models like (Soleimani et al., 2022), consumer healthcare questions ((Yadav et al., 2022); (Yadav and Caragea, 2022); (Yadav et al., 2023); (Savery et al., 2020)),

Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), pages 398–408

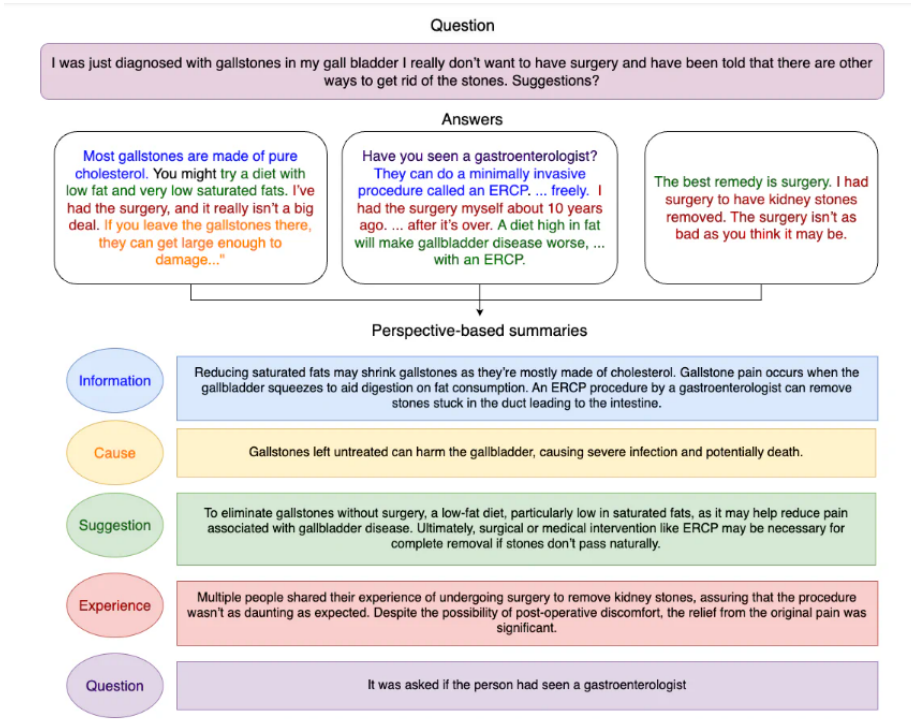

Figure 1: Task A: Span Prediction (highlighted spans), Task B: Summary Generation. (Source - (Agarwal et al., 2025))

and medical notes (Hsu et al., 2020).

(Fabbri et al., 2021) work on a QA dataset with sentence-level spans with query-focused multi-perspective abstractive summarization. (Joshi et al., 2020a) and (Michalopoulos et al., 2022) accomplish the same by exploiting local and global features of the text. CTRLsum (He et al., 2020) introduces a novel framework for controllable summarization that allows interaction during inference through textual input. CQASumm (Chowdhury and Chakraborty, 2018) highlight the issues with high-variance, opinion-based CQA data often having contradicting opinion and the challenges of applying Multi document summarization (MDS) on it.

In AnswerSumm (Fabbri et al., 2022), they use a model to extract sentences similar to the query. SpanBERT (Joshi et al., 2020b) extends BERT with a pre-training method, to better represent and predict spans of text. (Abaho et al., 2021) use both word-level and sentence-level attention to jointly perform span detection and outcome classification in the medical domain.

In this task the spans need not be complete sentences but rather can be phrases as well. The organizers of this task have annotated the dataset and proposed a prompt-driven control-label summarization model for the same.

3 Dataset

The dataset (Naik et al., 2024) used for the PerAnsSumm 2025 shared task consists of health-related questions and user-generated answers annotated with perspective categories. Each sample is a community Question-Answer thread (CQA) which includes a health-related question, an optional context providing additional background information, and a set of user answers. Specific spans within the answers are labeled according to one of five perspectives: Cause, Suggestion, Experience, Question, and Information. Additionally, each sample includes summaries that concisely represent the extracted spans for each perspective.

3.1 Dataset Statistics

The dataset is divided into training and validation sets, comprising 2,236 and 959 samples, respectively. During our Exploratory Data Analysis (EDA), we found that 4 samples in the training set and 3 samples in the validation set were incorrectly annotated. The spans in these samples were selected from the user context instead of the user answers, which goes against the task instructions. As a result, we discarded these samples, leaving us with 2,232 training samples and 956 validation samples. Among the validation samples, we randomly selected 300 samples as a test set to evaluate both open-source LLMs and proprietary models. The remaining 656 samples were used as a validation set for fine-tuning open-source LLMs.

Context availability varies, with 821 training samples containing context and 1,415 without it, while in the validation set, 350 samples include context and 606 do not include context.

The distribution of perspective categories reveals that Information and Suggestion are the most prevalent, whereas Cause and Question are less frequent. The complete label distribution across training and validation sets is illustrated in Figure 2.

A similar trend is observed in span counts, where Information spans appear most frequently, followed by Suggestion, Experience, Cause, and Question. The full span distribution can be seen in Figure 3.

4 Experimentations

4.1 Span Prediction

Span prediction involves identifying and classifying relevant spans within user responses based on predefined perspective categories. The models were evaluated using multiple performance metrics such as Classification Macro F1, Classification Weighted F1, Strict Matching Precision, Strict Matching Recall, Strict Matching F1, Proportional Matching Precision, Proportional Matching Recall, and Proportional Matching F1, ensuring a comprehensive assessment of both classification accuracy and span alignment.

4.1.1 LLM Fine-tuning

To effectively predict spans corresponding to different perspectives, we fine-tuned multiple open-source large language models, including Llama-3.1 8B (base model), Llama-3.2 3B (base model), and Gemma-2 9B (4-bit quantized model). The models were trained on the training set with Unsloth (Daniel Han and team, 2023) using zero-shot fine-tuning for 3 epochs with a learning rate of 2e-4 and validated on the validation set. The models were evaluated on the test set.

Among all models, the Llama-3.1 8B (base model) achieved the highest scores in classification, with a Classification Macro F1 of 0.7890, Classification Weighted F1 of 0.8360, and Strict Matching F1 of 0.2421. Meanwhile, the Gemma-2 9B (4-bit quantized model) outperformed others in proportional matching, achieving the highest Proportional Matching F1 score of 0.6652. A detailed comparison of these results is presented in Table 1.

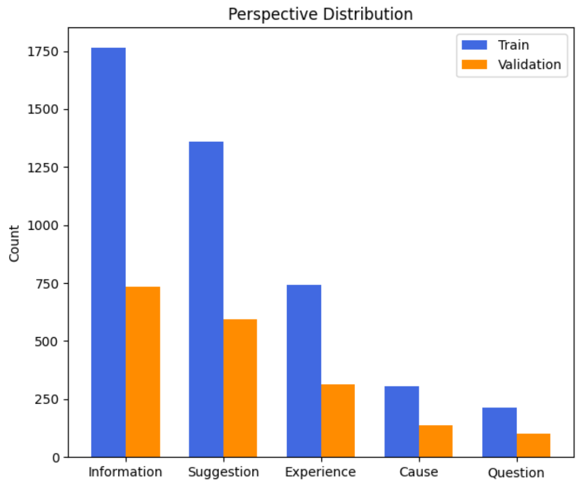

Figure 2: This figure shows the distribution of perspective categories in the training and validation datasets.

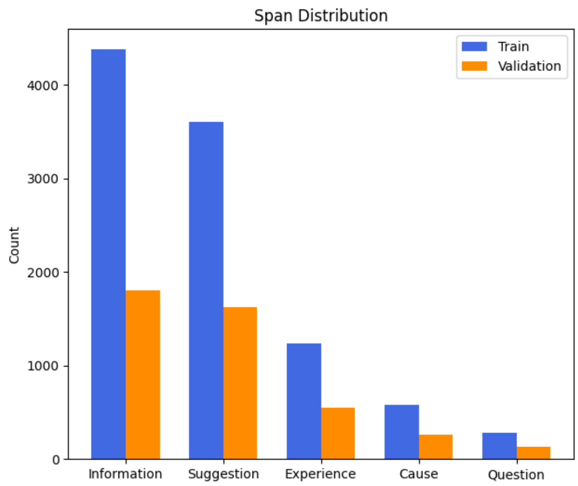

Figure 3: This figure shows the distribution of spans across perspective categories in the training and validation datasets. Each perspective category may contain one or more spans.

4.1.2 Proprietary Models

In addition to fine-tuning open-source models, we experimented with proprietary models, including GPT-4o, o1, Gemini-1.5 Pro, and Gemini-2 Flash Experimental. These models were evaluated using few-shot prompting, where we provided eight examples as context. We carefully selected these eight examples to mirror the label distribution in the training set. Two examples contained only one perspective, while one example included all five perspectives. The remaining examples featured

Metric	L3.1-8B	L3.2-3B	G2-9B (4b)	o1	o1 (50)	FL	FL (50)	o1 (CoT)	4o	Pro
C M F1	**0.7890**	0.6759	0.7102	0.7624	0.7601	0.7317	0.7102	0.7760	0.6770	0.7279
C W F1	0.8360	0.7545	0.8135	0.8404	0.8315	0.8305	0.8213	**0.8464**	0.7443	0.8258
S M P	**0.2734**	0.0958	0.0972	0.0611	0.0553	0.0627	0.0616	0.0432	0.0506	0.0618
S M R	**0.2172**	0.0758	0.0961	0.1114	0.0657	0.1118	0.1097	0.0568	0.0613	0.1089
S M F1	**0.2421**	0.0846	0.0967	0.0789	0.0601	0.0804	0.0789	0.0491	0.0554	0.0789
P M P	**0.7384**	0.6623	0.6479	0.6150	0.5903	0.6856	0.6759	0.6030	0.6615	0.6856
P M R	0.5436	0.5012	**0.6833**	0.6582	0.5358	0.6405	0.6674	0.5117	0.4474	0.6727
P M F1	0.6262	0.5706	0.6652	0.6359	0.5617	0.6623	0.6716	0.5536	0.5338	**0.6791**

Table 1: Performance comparison of various open-sourced and proprietary large language models for the span prediction task on the 300-sample holdout test set. **C M F1** and **C W F1** correspond to **Classification Macro F1** and **Classification Weighted F1**. **S M P**, **S M R**, and **S M F1** correspond to **Strict Matching Precision**, **Strict Matching Recall**, and **Strict Matching F1-score**. **P M P**, **P M R**, and **P M F1** correspond to **Proportional Matching Precision**, **Proportional Matching Recall**, and **Proportional Matching F1-score**. **L3.1-8B**, **L3.2-3B**, **G2-9B (4b)**, **o1 (50)**, **FL (50)**, **o1 (CoT)**, **4o**, and **Pro** represent **Llama-3.1 8B**, **Llama-3.2 3B**, **Gemma-2 9B (4-bit)**, **o1 (50-shot)**, **Gemini-2 Flash Experimental (50-shot)**, **o1 (Chain-of-Thought Prompting)**, **GPT-4o**, and **Gemini-1.5 Pro** respectively.

Metric	G2-9B (4b)	L3.1-8	4o	o1	o1 (CoT)	Pro	FL
Rouge-1	**0.5457**	0.4812	0.4911	0.4976	0.3380	0.5020	0.5323
Rouge-2	**0.2861**	0.2218	0.2337	0.2292	0.1160	0.2339	0.2713
Rouge-L	**0.4909**	0.4187	0.4211	0.4239	0.2810	0.4424	0.4765
BERTScore	0.9099	0.8611	0.8714	0.8972	0.8230	0.9064	**0.9103**
METEOR	**0.4754**	0.4529	0.4227	0.4176	0.2530	0.4154	0.4494
BLEU	**0.2137**	0.1923	0.1691	0.1992	0.0570	0.1792	0.2018

Table 2: Performance comparison of various open-sourced and proprietary large language models for the summarization task on the 300-sample holdout test set.

two, three, or four perspectives. The evaluation was conducted on the test set.

Among all proprietary models, o1 with Chain-of-Thought (CoT) prompting gave us the best classification results among all proprietary models. Gemini-2 Flash Experimental performed best in Strict Matching F1, while Gemini-1.5 Pro achieved the highest Proportional Matching F1.

To assess the impact of increasing the number of examples in few-shot prompting, we conducted an additional experiment by increasing the number of examples from 8 to 50, selected using random sampling for o1 and Gemini-2 Flash Experimental. The results showed that providing more examples did not improve performance. In fact, for o1, the Strict Matching F1 decreased from 0.0921 (8 examples) to 0.0601 (50 examples), and the Proportional Matching F1 dropped from 0.6359 to 0.5617. Similarly, for Gemini-2 Flash Experimental, the Classification Macro F1 declined from 0.7317 to 0.7102, and the Classification Weighted F1 decreased from 0.8305 to 0.8213. Although Strict Matching F1 and Proportional Matching F1 showed slight improvements, the gains were marginal. A detailed comparison of all the experiments is presented in Table 1.

4.2 Summarization

Once the relevant spans were identified for each perspective category, the next step was to generate a summary that effectively captured the key information from those spans. The models were evaluated using standard metrics such as ROUGE-1, ROUGE-2, ROUGE-L, BERTScore, METEOR, and BLEU.

4.2.1 LLM Fine-tuning

We fine-tuned Gemma-2 9B (4-bit quantized model) and Llama-3.1 8B (base model) to generate

Metric	S1	S2	S3	S4	S5	S6	S7	S8	S9
A + B	0.3964	0.4427	0.3940	0.4440	0.4083	**0.4495**	0.3833	0.4467	0.4407
C M F1	0.8628	0.7933	0.8656	0.8509	0.8581	**0.8656**	0.7849	0.8656	0.8656
C W F1	0.9092	0.8634	0.9140	0.8992	0.8900	**0.9140**	0.8396	0.9140	0.9140
S M P	0.1352	0.1768	0.1491	**0.1775**	0.1748	0.1765	0.1552	0.1765	0.1765
S M R	0.1257	0.2667	0.1562	0.2705	0.1162	**0.2743**	0.1200	0.2743	0.2743
S M F1	0.1303	0.2126	0.1526	0.2143	0.1396	**0.2148**	0.1353	0.2148	0.2148
P M P	0.5189	**0.6793**	0.5892	0.6641	0.5275	0.6597	0.4420	0.6597	0.6597
P M R	0.6857	**0.7396**	0.5648	0.7076	0.6350	0.7159	0.6145	0.7159	0.7159
P M F1	0.5907	**0.7081**	0.5767	0.6852	0.5763	0.6866	0.5142	0.6866	0.6866
A	0.5434	0.5947	0.5478	0.5996	0.5353	**0.6052**	0.4964	0.6052	0.6052
ROUGE-1	0.3580	0.4129	0.3407	0.4201	0.3533	**0.4345**	0.3318	0.4243	0.4048
ROUGE-2	0.1432	0.1818	0.1058	0.1812	0.1574	**0.1869**	0.1434	0.1753	0.1542
ROUGE-L	0.3210	0.3714	0.2881	0.3763	0.3184	**0.3878**	0.3017	0.3765	0.3510
BERTScore	0.8038	0.8048	0.8531	0.8318	0.7385	**0.8658**	0.7220	0.8621	0.8584
METEOR	0.3226	0.3713	0.2572	0.3719	0.3190	**0.3844**	0.3041	0.3509	0.3474
BLEU	0.0971	**0.1189**	0.0602	0.1127	0.1088	0.1124	0.0959	0.1134	0.1047
B_Relevance	0.3409	0.3768	0.3175	0.3823	0.3326	**0.3953**	0.3165	0.3838	0.3701
AlignScore	0.3665	**0.4458**	0.4043	0.4307	0.4359	0.4260	0.3991	0.4308	0.4369
SummaC	0.2433	0.2671	0.2291	0.2696	**0.2785**	0.2701	0.2750	0.2715	0.2570
B_Factuality	0.3049	0.3565	0.3167	0.3502	**0.3572**	0.3480	0.3370	0.3512	0.3470

Table 3: Performance comparison across all submissions evaluated on the provided 50 samples.

summaries from the predicted spans. Both models were trained on the training set with Unsloth (Daniel Han and team, 2023) using zero-shot fine-tuning for 3 epochs with a learning rate of 2e-4, validated on the validation set, and evaluated on the test set.

Among these two, Gemma-2 9B (4-bit quantized model) consistently outperformed the Llama-3.1 8B model across all evaluation metrics. A detailed comparison of the results is presented in Table 2.

4.2.2 Proprietary Models

In addition to fine-tuned models, we explored proprietary models, including GPT-4o, o1, Gemini-1.5 Pro, and Gemini-2 Flash Experimental, using a few-shot prompting approach with 8 examples. We used the same examples which were used the span prediction task. These models were evaluated on the test set. Among these models, Gemini-2 Flash Experimental consistently achieved the highest scores across all evaluation metrics. A detailed comparison of the results is presented in Table 2.

5 Submissions

During the competition's evaluation phase, we were given 50 test samples and made a total of nine submissions, each exploring different model configurations and techniques.

In our first submission, we fine-tuned the Gemma-2 9B (4-bit quantized) model on the training data and validated it on the validation data for span prediction and summarization. The second submission (S2) used Gemini-2 Flash Experimental, a proprietary model, for both tasks. The third submission (S3) introduced o1 with Chain-of-Thought (CoT) prompting to enhance reasoning capabilities.

In the fourth submission (S4), we used o1 (CoT) for classification and Gemini-2 Flash Experimental for span extraction and summarization. However, Gemini-2 Flash Experimental did not always adhere to the class predictions from o1, leading to inconsistencies in output. For the fifth submission (S5), we fine-tuned Gemma-2 9B (4-bit quantized) using a combined training and validation set.

Our sixth submission (S6) achieved the best over-

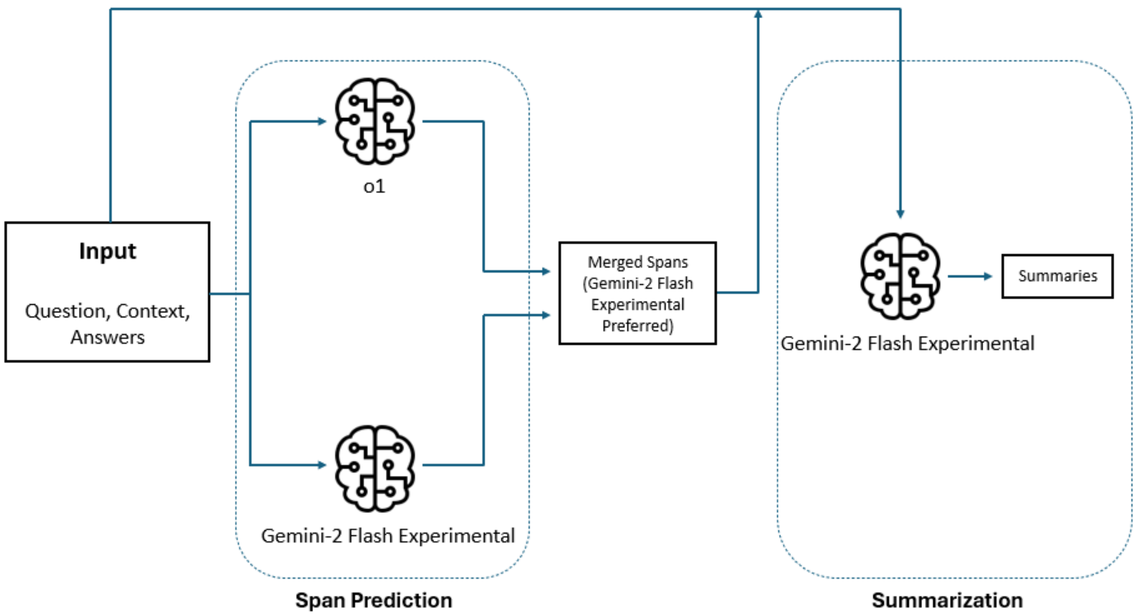

Figure 4: This figure illustrates the workflow of our best submission.

all performance. Here, we used o1 and Gemini-2 Flash Experimental for span extraction, ensuring that all classes predicted by o1 had corresponding spans. We noticed that Gemini's perspective classification was a proper subset of o1's. If Gemini-2 Flash Experimental did not generate spans for a perspective category but o1 did, we retained those from o1. When both models provided spans for a particular perspective, we used those from Gemini-2 Flash Experimental and discarded o1's. The final set of spans was then passed to Gemini-2 Flash Experimental for summarization. This submission achieved the highest Task A+B average score of 0.4495. The complete workflow is illustrated in Figure 4.

While evaluating the test data, we observed that all 50 samples included context, whereas two-thirds of the training data lacked it. To account for this, our seventh submission (S7) fine-tuned Gemma-2 9B using only samples that contained context. In the eigth submission (S8), we used o1 for classification, Gemini-2 Flash Experimental for span extraction, and increased the few-shot prompting examples from 8 to 16 to enhance summarization performance.

For our final submission (S9), o1 was used for span extraction, and Gemini-1.5 Pro was used for summarization. A detailed breakdown of the scores for all submissions is provided in Table 3.

In Table 3, the metric (A+B) denotes the combined average score of Task A and B, and (A) represents the score for Task A. The metrics (B_Relevance) and (B_Factuality) correspond to the relevance and factuality scores for Task B, respectively. AlignScore (Zha et al., 2023) and SummaC (Laban et al., 2022) are factual consistency evaluation metrics, designed to assess the alignment of generated summaries with the source text.

6 Discussion

In the final submissions we notice that o1 CoT performs well on the classification task (to predict perspectives present in user answers) as seen in Table 3. This is in line with our evaluations on the test set as well, where the classification weighted F1 of o1 CoT was the best as seen in Table 1. For the span extraction task, finetuned open-source models were performing on par with proprietary ones like Gemini-2 Flash Experimental and 1.5 Pro as seen in Table 1. For summarization Gemma-2 9B (4 bit) beats all other models as seen in Table 2. This demonstrates the efficiency of finetuning Large Language Models on downsteam tasks where even smaller models (less than 10 B parameters) can compete with and beat larger models like GPT 4o etc.

However, in the final submissions we see a large gap between open-sourced models like Gemma-2 9B (4-bit) (Submision 1) and proprietary models like Gemini-2 Flash Experimental (Submission 2) as seen in Table 3. The reason for such discrepancy can be due to difference in data distribution of the training and validation set released earlier and the final evaluation set of 50 samples on which the submissions were scored. One difference highlighted earlier was that the final evaluation set had the optional context section for all samples whereas, the training and validation set had approximately two-thirds of the samples without the context section. Another reason could be an inherent bias due to a small set of just 50 samples.

7 Conclusion

We test multiple open-source and proprietary LLM's for the task. Finetuning open-source smaller models like Llama 8b, 3b and Gemma 9b models yielded significant improvements from their base variants and even outperformed GPT 4o. This is likely because learning is significantly higher from finetuning when compared to in- context Learning with few shot examples. It is also difficult to capture all the details of the data in the few shot examples which is another reason why finetuning performs better. In our experiments, we observed that increasing the number of few-shot examples did not enhance performance. Hence finetuning is the better alternative.

Regardless, few proprietary LLM's particularly Gemini-2 Flash Experimental was able to beat the finetuned smaller models like Llama and Gemma on the final evaluation set of 50 samples on which submissions were scored. Possible reasons for a significant drop in performance during the final evaluation is discussed in the Discussions section. We also try a CoT prompt with o1 to accomplish both tasks in one go. We notice that the classification (perspective prediction) of o1 CoT is the best of all submissions (Table 3) which is largely in line with our experimentations (Table 1), but the spans and summaries of Gemini-2 Flash Experimental is better. Hence, we merge the spans of both models and choose Gemini's spans wherever possible. For perspectives where Gemini does not generate any spans but o1 does, we go ahead with the spans from o1. This ensures we utilize the better classification performance of o1 and use Gemini's span and summarization.

8 Limitations

The experiments carried out were mainly on a few selected open source and proprietary models. There are a number of open-sourced larger models which could have been finetuned for better performance. However, due to insufficient resources and time constraints we keep it as a possible future work. As for the proprietary models, more effort can be put in the prompting of these models. Things like a greater number of few shot prompts, different few shot examples can be tried. An ensemble approach using o1 and Gemini-2 Flash Experimental for span prediction, combined with the Gemma-2 9B model for summarization, could also be explored for improved performance.

9 Ethical Consideratons

Given that our dataset is from the medical and healthcare domain we take additional effort to comply with all ethical guidelines. As per the shared tasks instructions we use this dataset strictly for the task experiments and have not leaked this data to any third party. Since the data contains answers from multiple users there are some personal identification information like email addresses, website links etc. We make no effort to make contact or connect to these users on their social media handles. Also, we have cited all intellectual artifacts and resources to the best of our knowledge, ensuring proper attribution and adherence to ethical research practices.

References

Micheal Abaho, Danushka Bollegala, Paula Williamson, and Susanna Dodd. 2021. Detect and classify – joint span detection and classification for health outcomes. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8709–8721, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Tanya Chowdhury and Tanmoy Chakraborty. 2018. Cqasumm: Building references for community question answering summarization corpora. *Preprint*, arXiv:1811.04884.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of* *the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexander R. Fabbri, Xiaojian Wu, Srini Iyer, and Mona Diab. 2021. Multi-perspective abstractive answer summarization. *Preprint*, arXiv:2104.08536.

Alexander R. Fabbri, Xiaojian Wu, Srini Iyer, Haoran Li, and Mona Diab. 2022. Answersumm: A manually-curated dataset and pipeline for answer summarization. *Preprint*, arXiv:2111.06474.

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. *Preprint*, arXiv:2407.21783.

Junxian He, Wojciech Kryściński, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2020. Ctrlsum: Towards generic controllable text summarization. *Preprint*, arXiv:2012.04281.

Chao-Chun Hsu, Shantanu Karnwal, Sendhil Mullainathan, Ziad Obermeyer, and Chenhao Tan. 2020. Characterizing the value of information in medical notes. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2062–2072, Online. Association for Computational Linguistics.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *Preprint*, arXiv:1904.05342.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020a. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3755–3763, Online. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020b. Spanbert: Improving pre-training by representing and predicting spans. *Preprint*, arXiv:1907.10529.

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.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Preprint*, arXiv:1910.13461.

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. *Preprint*, arXiv:1907.11692.

George Michalopoulos, Kyle Williams, Gagandeep Singh, and Thomas Lin. 2022. MedicalSum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4741–4749, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and

Yury Malkov. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Max Savery, Asma Ben Abacha, Soumya Gayen, and Dina Demner-Fushman. 2020. Question-driven summarization of answers to consumer health questions. *Preprint*, arXiv:2005.09067.

Amir Soleimani, Vassilina Nikoulina, Benoit Favre, and Salah Ait Mokhtar. 2022. Zero-shot aspect-based scientific document summarization using self-supervised pre-training. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 49–62, Dublin, Ireland. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Shweta Yadav and Cornelia Caragea. 2022. Towards summarizing healthcare questions in low-resource setting. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2892–2905, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Shweta Yadav, Ștefan Cobeli, and Cornelia Caragea. 2023. Towards understanding consumer healthcare questions on the web with semantically enhanced contrastive learning. pages 1773–1783.

Shweta Yadav, Deepak Gupta, and Dina Demner-Fushman. 2022. Chq-summ: A dataset for consumer healthcare question summarization. *Preprint*, arXiv:2206.06581.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. Biobart: Pretraining and evaluation of a biomedical generative language model. *Preprint*, arXiv:2204.03905.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. *Preprint*, arXiv:2305.16739.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *Preprint*, arXiv:1912.08777.