

On The Persona-based Summarization of Domain-Specific Documents

¹Ankan Mullick ¹Sombit Bose* ¹Rounak Saha* ²Ayan Kumar Bhowmick

¹Pawan Goyal ¹Niloy Ganguly ²Prasenjit Dey ²Ravi Kokku

{ankanm, sbcs.sombit, runk20}@kgpian.iitkgp.ac.in

{pawang, niloy}@cse.iitkgp.ac.in, {ayan, prasenjit, ravi}@merlyn.org

¹Computer Science and Engineering Department, IIT Kharagpur, India. ²Emergence AI

Abstract

In an ever-expanding world of domain-specific knowledge, the increasing complexity of consuming, and storing information necessitates the generation of summaries from large information repositories. However, every persona of a domain has different requirements of information and hence their summarization. For example, in the healthcare domain, a persona-based (such as Doctor, Nurse, Patient etc.) approach is imperative to deliver targeted medical information efficiently. Persona-based summarization of domain-specific information by humans is a high cognitive load task and is generally not preferred. The summaries generated by two different humans have high variability and do not scale in cost and subject matter expertise as domains and personas grow. Further, AI-generated summaries using generic Large Language Models (LLMs) may not necessarily offer satisfactory accuracy for different domains unless they have been specifically trained on domain-specific data and can also be very expensive to use in day-to-day operations. Our contribution in this paper is two-fold: 1) We present an approach to efficiently fine-tune a domain-specific small foundation LLM using a healthcare corpus and also show that we can effectively evaluate the summarization quality using AI-based critiquing. 2) We further show that AI-based critiquing has good concordance with Human-based critiquing of the summaries. Hence, such AI-based pipelines to generate domain-specific persona-based summaries can be easily scaled to other domains such as legal, enterprise documents, education etc. in a very efficient and cost-effective manner.

1 Introduction

In the rapidly expanding digital world, the exponential growth of domain-specific knowledge has posed unprecedented challenges in efficiently storing and consuming vast information repositories.

With the increasing complexity of managing such information, the need for generation of precise and specific summaries becomes important. This need becomes particularly evident for domain-specific data as there exist different personas within a domain who have different information requirements which should be reflected in generated summaries. For instance, if we consider the healthcare domain, there exist diverse personas¹ ranging from healthcare professionals like doctors and nurses to patients who require targeted information customized to their specific roles and comprehension levels.

Traditional generic approaches to summarization have often relied on humans to perform this high cognitive load task. However, as the volume and diversity of information burgeon with growing number of domains and personas, human generated persona-based summaries encounter limitations in scalability, cost-effectiveness, and consistency. The inherent subjectivity and variability among different human summarizers hinder the reliability and efficiency of such an approach. Although there have been several approaches in prior literature with focus on generic summaries through extractive and abstractive methods (Paulus et al., 2017; Erkan and Radev, 2004) as well as goal-oriented summaries (Hayashi et al., 2021; Zhu et al., 2022) but none of them have focused on persona-based summarization of domain-specific information. Goldsack et al. (2023); Luo et al. (2022) focus on building layman-summarization comprehensible to non-technical audiences but do not differentiate the various technical summaries based on persona and they also do not use LLMs as an alternative evaluator. Our work differs in the sense that we develop a pipelined approach that generates persona-specific training summaries (doctor, patient, normal person), fine-tune small-size LLMs on this data, and use GPT-4 to efficiently evaluate summary quality.

* Authors contributed equally

¹ <http://tiny.cc/x1guwz>

One possible solution is to harness the power of generic large language models (LLMs) such as GPT-4 to automate the generation of persona-based summaries as such models have been used to generate data for other NLP tasks (Sun et al., 2023; Yu et al., 2023). ChatGPT² is also used in educational data generation (Kieser et al., 2023; Maddigan and Susnjak, 2023). However, AI-generated summaries using generic LLMs may not be guaranteed to achieve optimal accuracy across different domains unless they are trained on domain-specific data and they can also be very expensive to use for daily repeated inferences. In this paper, we take a step towards introducing a two-fold contribution aimed at overcoming these challenges of generating domain-specific persona-based summaries. Firstly, we present an efficient approach towards the training of domain-specific, small-sized Large Language Models (LLMs) on a corpus related to healthcare domain. Though data distillation from a stronger model for supervised fine-tuning is a standard method, the novelty of our work lies in the fact that we effectively employ the approach in this context to build a cost-optimised summarization framework catering to different healthcare persona given the scarcity of domain-specific data. This approach addresses the limitations of generic LLMs by aligning the trained model specifically to the intricacies of summaries in the healthcare domain. Moreover, we showcase the effectiveness of utilizing AI-based critiquing for the evaluation of summarization quality, providing a more automated and scalable solution.

Secondly, we demonstrate the strong agreement between AI-based and human-based critiquing of generated summaries, establishing the reliability of our proposed approach. This not only validates the effectiveness of domain-specific small LLM-based models in generating accurate summaries but also opens up avenues for scalability across diverse domains. The implications of our findings extend beyond healthcare, as the proposed AI pipeline can be seamlessly adapted to other domains, including legal, corporate documents, education, and more, offering a versatile and cost-effective solution for generating persona-based summaries.

2 Proposed Framework

We describe here our framework for training our small-size domain-specific LLM, generation of

² <https://chat.openai.com/>

| Filtering Steps with Criteria (removed) | % |
|---|------|
| Step 1: Too many special characters and other string (HTML tags and #) | 1.52 |
| Step 2: Incomplete Summary (By checking punctuations) | 0.86 |
| Step 3: Conflict identification - Very similar summaries of different persona | 1.12 |
| Step 4: If the summary contains Medical Terms or numbers not present in the document (using QuickUMLS - https://github.com/Georgetown-IR-Lab/QuickUMLS/) | 1.39 |
| Overall summaries filtered out | 4.89 |

Table 1: Step-by-Step Data Filtering data for finetuning and evaluation, and other model baselines that we compare our model against.

2.1 Dataset

We create persona-based dataset (named ‘*Persona-Data*’) utilizing GPT-4³ with specific prompts on 1455 articles from the publicly available WebMD⁴ website. The mean ratio between summary length and document length is 0.2 : 1. After data generation using GPT-4, we did a step by step validation of the generated summaries using an automated approach followed by manual verification to filter out the bad generations. Around 4.89% of document-summary pairs were filtered out. We provide the detailed filtering steps with criteria and removal pairs (in %) as shown in Table 1.

These articles, related to healthcare, form the basis for creation of their summaries related to three distinct persona: (a) *Doctor*: Summaries focus on medical terminology, guidelines and provide detailed technical information suitable for medical professionals. (b) *Patient*: Summaries are easily understandable, addressing patient concerns without excessive technical jargon, focusing on top-level information. (c) *Normal Person*: Summaries are tailored for a general audience without medical background, presented in simple language and engaging for laypersons while avoiding technical terms. The dataset comprises 1091, 73 and 291 articles for training, validation, and testing respectively. Additionally, we select 50 WebMD articles and generate manual summaries for three personas (termed as *Annotated-Data*) using the Prolific⁵ annotation platform and Doctors to evaluate the GPT-4 generated summaries against human curated summaries⁶ (Details are in Appendix F).

2.2 Model Architecture

Our training process consists of employing small foundation LLMs such as Llama2 and finetuning such models on the training set of the WebMD data

³ <https://openai.com/gpt-4>

⁴ <https://www.webmd.com/>

⁵ <https://www.prolific.com/>

⁶ Code/Dataset details are in <https://github.com/ankan2/persona-healthcare>

| Model | Rouge1 | Rouge2 | RougeL | Meteor | Bleu | BERT-Prec | BERT-Rec | BERT-F1 |
|----------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Falcon | 23.3 | 5.0 | 12.8 | 12.4 | 1.3 | 78.6 | 73.2 | 75.8 |
| BART | 23.6 | 8.1 | 15.0 | 11.3 | 0.8 | 83.8 | 74.9 | 79.1 |
| Pegasus | 27.8 | 8.3 | 16.2 | 14.3 | 2.0 | 80.7 | 75.0 | 77.8 |
| T5-FT | 42.6 | 15.6 | 23.7 | 30.2 | 11.4 | 84.2 | 80.7 | 82.4 |
| FT5-FT | 41.2 | 15.5 | 23.3 | 29.0 | 11.2 | 83.9 | 79.9 | 81.9 |
| LED-B | 33.3 | 9.3 | 17.4 | 17.3 | 3.3 | 80.7 | 76.7 | 78.6 |
| LED-L | 39.8 | 12.3 | 25.6 | 25.3 | 9.3 | 82.5 | 78.7 | 80.6 |
| L-V-7b | 14.2 | 4.0 | 8.5 | 6.7 | 0.6 | 72.0 | 65.7 | 68.7 |
| L-V-13b | 24.1 | 7.3 | 14.3 | 11.5 | 1.1 | 80.8 | 73.5 | 77.0 |
| L-V-70b | 25.1 | 7.2 | 13.7 | 16.2 | 3.7 | 65.7 | 64.6 | 65.1 |
| L-F-7b | 45.9 | 17.2 | 25.2 | 32.6 | 12.5 | 83.6 | 82.6 | 83.1 |
| L-F-13b | 53.7 | 24.3 | 33.8 | 38.3 | 18.4 | 87.4 | 85.3 | 86.3 |

Table 2: Traditional Metrics’ based evaluation results (all are in %)

described above. Specifically, we use supervised fine-tuning approach (Li et al., 2023) on the pre-trained vanilla model versions of Llama2.

Llama2⁷: We perform supervised fine-tuning (SFT) on the pretrained vanilla Llama2-7b and Llama2-13b models using the training data that consists of prompt-completion pairs where the prompt comprises a WebMD article in the training set and the instruction to generate a summary based on a persona (doctor/patient/normal person) and the completion is the corresponding persona-based summary generated using GPT-4. We use a parameter efficient finetuning approach i.e. Quantized Low-Rank Adaptation (QLoRA) (Detrmers et al., 2023) to optimize the training process. After training, the finetuned Llama2-7b and Llama2-13b models (referred to as L-F-7b and L-F-13b respectively) acquire the ability to generate a persona-based summary for a given medical article depending on the persona specified in the prompt.

Baselines: For comparison with our finetuned models on the persona-based summary generation task, we use different state-of-the-art models as baselines such as Falcon 7b-instruction tuned model (Penedo et al., 2023), BART-large (Lewis et al., 2019), instruction-tuned Pegasus (Zhang et al., 2020) and Longformer (Beltagy et al., 2020) Base (LED-B), Large (LED-L)⁸. Besides these, we also use finetuned versions of T5-Large (Raffel et al., 2020) (T5-FT) and Flan-T5-Large (Chung et al., 2022)

(FT5-FT) on our training data as baselines. Further, we also compare the performance with the different vanilla Llama2 model variants (7b, 13b, 70b referred to as L-V-7b, L-V-13b and L-V-70b respectively).

3 Evaluation and Results

We evaluate the performance of our finetuned models (L-F-7b and L-F-13b) in terms of generating high quality persona-based summaries for medical articles and compare against the baseline models.

Evaluation metrics: Our evaluation relies on two different approaches:

(i) *Traditional* - Here we use traditional metrics such as Rouge [1, 2 and L] (Lin, 2004), Meteor (Banerjee and Lavie, 2005), Bleu (Papineni et al., 2002), BERTScore (Zhang et al., 2019) [Precision (BERT-Prec), Recall (BERT-Rec) and F1-score (BERT-F1)] to assess the quality of generated summaries.

(ii) *GPT-4 critique* - Here we use the GPT-4 LLM as a critic to evaluate the quality of the model generated summaries against the gold standard GPT-4 generated summary (Section 2) from different dimensions. Specifically, we provide suitable critique based prompts to GPT-4 where we evaluate the summaries based on a set of five predefined criterias (termed as *GPT-4 criteria*) defined below: *Criteria 1: Relevance (Rel)*: The extent to which the generated persona-based summary is relevant to the intended persona (doctor/patient/normal person) given the document.

Criteria 2: Coverage (Cov): The extent to which

⁷ <https://ai.meta.com/llama/>

⁸ <https://huggingface.co/allenai/led-base-16384>

| Model | Rouge1 | Rouge2 | RougeL | Meteor | Bleu | BERT-Prec | BERT-Rec | BERT-F1 |
|----------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Doctor | 53.9 | 24.7 | 34.0 | 37.0 | 18.7 | 88.0 | 85.1 | 86.5 |
| Patient | 53.5 | 24.2 | 33.5 | 36.7 | 18.3 | 87.2 | 84.8 | 86.0 |
| Nor-Per | 53.6 | 23.9 | 33.9 | 41.0 | 18.1 | 86.9 | 86.1 | 86.5 |
| Average | 53.7 | 24.3 | 33.8 | 38.3 | 18.4 | 87.4 | 85.3 | 86.3 |

Table 3: Traditional Metrics’ based evaluation on different persona using Llama2-13b Finetuning model (in %)

the generated persona-based summary correctly covers important key points described in the gold standard persona-based summary of the document.

Criteria 3: Impurity (Imp): The extent to which the persona-based summary does not contain information specific to all other possible personas $\{persona_set - persona\}$.

Criteria 4: Quality (Qlt): The extent to which the persona-based summary is of overall good quality from the perspective of the intended persona.

Criteria 5: Goodness (Gds): Extending from 4, we manually verify the goodness of the summary.

(Details of these criteria with prompts are provided in Appendix D).

Results: We provide a comparison of our finetuned models against the baselines in terms of both traditional metrics as well as GPT-4 criteria in Tables 2 and 4 respectively (all values are in %) on the WebMD test set of size 873 (prompt specific to three different persona each for 291 articles). Table 2 infers that both our finetuned models (L-F-7b and L-F-13b) achieve superior performance compared to the baseline methods in terms of traditional metrics. In fact, finetuned Llama2-13b (L-F-13b) outperforms the baselines in terms of all the traditional metrics, demonstrating the superiority of our finetuning approach which helps to adapt the model to the healthcare domain and perform better on specific applications such as persona-based summarization. Similar observation holds true when we compare the values of the GPT-4 critique based criteria shown in Table 4. Here we also compare the quality of the gold standard GPT-4 generated summaries in terms of the GPT-4 critique based criteria. We find that the finetuned Llama2-13b model (L-F-13b) can generate summaries pretty close in quality to the gold standard, while being much faster in terms of training and inference time as well as cost-effective and cheaper in terms of memory requirement.

Framework: We use 80GB A100 GPU, 210MHz clock cycle along with NLTK/SpaCy python packages for all experiments. For 6 epochs, Llama2-13b takes 20 hrs for finetuning and 3 hrs for inference, Llama2-7b takes 8 hrs for finetuning and 2.5 hrs for inference (Details in Appendix E.3).

Ablation study: Here we investigate the quality of persona-based summaries generated by different variations of our best performing finetuned Llama2-13b (L-F-13b) model on WebMD test set:

(A) Performance specific to different persona: Table 3 and 5 shows the performance in terms of

| Model | Rel | Cov | Imp | Qlt | Gds |
|----------------|-------------|-------------|-------------|-------------|-------------|
| Falcon | 56.3 | 45.4 | 82.0 | 50.6 | 46.8 |
| BART | 65.4 | 42.6 | 84.8 | 49.7 | 25.8 |
| Pegasus | 47.7 | 33.0 | 74.0 | 36.1 | 11.5 |
| T5-FT | 72.4 | 70.1 | 84.9 | 67.6 | 78.1 |
| FT5-FT | 72.2 | 70.2 | 88.0 | 68.3 | 80.3 |
| LED-B | 36.1 | 19.2 | 79.3 | 31.3 | 17.6 |
| LED-L | 69.1 | 56.2 | 82.3 | 59.3 | 56.6 |
| L-V-7b | 19.1 | 18.4 | 41.3 | 16.6 | 15.2 |
| L-V-13b | 32.1 | 29.1 | 73.1 | 28.5 | 23.4 |
| L-V-70b | 49.7 | 45.1 | 78.0 | 46.4 | 47.8 |
| L-F-7b | 75.8 | 58.7 | 85.8 | 63.8 | 58.6 |
| L-F-13b | 93.5 | 90.1 | 91.7 | 88.5 | 99.1 |
| GPT-4 | 98.2 | 96.3 | 98.6 | 98.5 | 99.7 |

Table 4: GPT-4 Critique evaluation results (in %)

standard evaluation metrics and outcomes of GPT-4 critique based criteria for each of the three persona [Doctor, Patient and Normal Person (Nor-Per)] for the best performing Llama2-13b Finetuned model (L-F-13b). We observe that the model performs uniformly across the three persona which confirms that our finetuned model generalizes well across multiple persona, generating distinct persona-based summaries for the same medical article.

| Persona | Rel | Cov | Imp | Qlt | Gds |
|---------|------|------|------|------|------|
| Doctor | 90.0 | 89.1 | 91.0 | 86.2 | 98.8 |
| Patient | 94.4 | 90.4 | 92.2 | 88.5 | 99.0 |
| Nor-Per | 93.2 | 91.0 | 91.8 | 87.7 | 99.3 |
| Average | 93.5 | 90.1 | 91.7 | 88.5 | 99.1 |

Table 5: GPT-4 critique on different persona using Llama2-13b Finetuning model (in %)

(B) Validation of GPT-4 generated gold standard summaries: To verify the robustness of the GPT-4 generated summaries for the WebMD articles and to mitigate the GPT-4 introduced inherent bias in generated summaries, we perform different types of human annotation experiments:

(i) Persona-based Summary of GPT-4: We randomly select 50 different WebMD articles and provide three different persona-based (doctor, patient, normal person) summaries (without their actual labels) to three different doctors with domain knowledge expertise along with a good working proficiency in English and ask them to identify the intended persona, i.e. - which summary belongs to which specific persona. Initial human labeling is done by two doctors and any annotation discrepancy is checked and resolved by the third doctor after discussing with others. The inter-annotator agreement is found to be 0.91. On comparing with actual persona labels, we found that human labels have 86.67% accuracy for correctly identifying the actual persona which shows the reliability of GPT-4

generated persona-based summaries.

(ii) Content Quality Check: We ask human annotators (doctors) to annotate summaries on the basis of whether the persona-based summary is relevant while correctly covering appropriate key points based on information need of different persona and its overall usefulness. 96% of the GPT-4 generated summaries for different personas are found to be useful by human annotators (doctors).

(iii) GPT-4 Generated and Ground Truth Summary Check: Both doctors with domain expertise and GPT-4 evaluate 50 document-summary pairs in terms of whether the persona-based summary is relevant and correctly covers appropriate key points based on information need of different persona, each for GPT-4 generated summaries and ground truth summaries generated by annotators. We obtain an inter-annotator agreement of 0.893 which signifies strong consensus between human and GPT-4 based evaluation. We separately test our best fine-tuned (Llama2-13b-FT) model on the human-generated summaries (50 articles) and obtain the following scores for different metrics - Traditional: R1-52.9, R2-24.1, RL-33.2, Meteor-38.7, Bleu-18.0, Bert-P-87.7, Bert-R-84.9, Bert-F1-86.3; GPT-4 criteria: Rel-91.2, Cov-90.8, Imp-90.4, Qlt-88.7, Gds-98.5. This shows that there is strong alignment between results of on human generated and GPT-4 generated summaries, signifying the high quality of GPT-4 generated summaries.

(C) Validation of finetuned model generated summaries: To further investigate the reliability of our finetuned model generated summaries, we choose 50 different WebMD articles and provide persona-based summaries for each persona (generated by GPT-4 in ground truth v/s Llama2-13b finetuned model generated) to two doctors to annotate: (i) whether finetuned generated summary is better, (ii) Both are Good, (iii) Ground Truth/GPT-4 summary is better and (iv) Both are bad. We find that for - 20% cases Llama2-13b finetuned model summaries are better (i), for 50% cases both finetuned and ground truth generated summaries are good (ii) and rest 30% cases ground truth generated summaries are better (iii) and no instance is found where both performs bad (iv).

(D) Different LLM Evaluators: We evaluate the fine-tuned model generated summaries in the test set with Gemini model⁹ keeping the same prompts and criteria as used earlier for GPT-4 and the obtain

values of the same LLM based criteria are: Rel - 95.2, Cov - 92.4, Imp - 87.6, Qlt - 90.7, Gds - 99.4. Thus, Gemini scores are also aligned with GPT-4 scores with a correlation coefficient of 0.808 (Gemini provides higher scores for all criteria except Criteria 3 - 'Imp'). This verifies that the GPT-4 based evaluation is impartial and robust.

(E) Llama2-13b performance on Other data: We test our best performing Llama2-13b finetuned model on healthcare domain articles of OASUM (Yang et al., 2022) dataset which is publicly available. We select OASUM articles with aspects related to healthcare [Death, Diagnosis, Differential Diagnosis and Diagnosis Classification] and obtain 234 such documents. We perform GPT-critique based evaluation and observe that 82.77% of the summaries are labeled as good which signifies the robustness of our model in terms of generating high quality summaries.

4 Conclusion

In this paper, we propose a framework for the efficient training of a small foundation LLM on AI-generated datasets to obtain high quality domain-specific persona-based summaries. Our focus is on training a finetuned version of Llama2 on a corpus related to healthcare domain such that the trained model captures the intricacies of persona-based summaries in healthcare domain. We also demonstrate the effectiveness of using AI-based critiquing for the evaluation of the model generated summaries, providing a more automated and scalable solution. Our experiments also reveal the superior quality of persona-based summaries generated by our finetuned model compared to contemporary baselines. Further, AI-based critiquing of the summaries show high inter-annotator agreement with human-based critiquing methods, further confirming the effectiveness of our proposed approach. We plan to extend our work in generating accurate persona-based summaries for documents in other domains such as legal, enterprises, education and more, which is the focus of our future work.

Acknowledgements

The work was supported in part by a research grant from IIT KHARAGPUR AI4ICPS I HUB FOUNDATION.

⁹ <https://gemini.google.com/>

Limitation and Discussion

There are a few limitations in our works - (i) Not all LLMs are useful (similar to GPT-4) to generate personalized contents properly - like GPT2, GPT3.5, Llama2-vanilla models do not perform very well mostly due to hallucination and not covering important informations. (ii) We only explore the data from healthcare domain, but we plan to extend our work to other domains such as legal, corporate and education among others. (iii) Our experimental dataset is only English in healthcare domain, we wish to extend the work in multilingual setup, specifically for low-resource settings in diverse domains. (iv) Prompt command is very important. Unless, we specifically mention with explanation in Prompt about different *Persona* (Doctor/Patient/Normal Person), GPT-4 does not perform well to generate appropriate summary.

Ethical Concerns

We use the publicly available content of the WebMD platform for non-commercial and academic purpose only without violating any ethical concerns. The dataset neither reveals any personal sensitive information of the patients nor any toxic statement. Consent has been taken from all annotators including doctors. For experiments, we use publicly available free frameworks - Llama2, Falcon, BART, Pegasus, T5-FT, Flan-T5 (FT5-FT), LED-Base, LED-Large, Llama2-7b, 13b and 70b - vanilla and finetune.

References

- Nadeem Akhtar, Nashez Zubair, Abhishek Kumar, and Tameem Ahmad. 2017. Aspect based sentiment oriented summarization of hotel reviews. *Procedia computer science*, 115:563–571.
- 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.
- Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.
- Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.
- Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021. Medically aware gpt-3 as a data generator for medical dialogue summarization. In *Machine Learning for Healthcare Conference*, pages 354–372. PMLR.
- Sumit Chopra, Michael Auli, and Alexander M Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 93–98.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.
- Maximin Coavoux, Hady Elsahar, and Matthias Gallé. 2019. Unsupervised aspect-based multi-document abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 42–47.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.
- Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.
- Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2024. Llm-based nlg evaluation: Current status and challenges. *arXiv preprint arXiv:2402.01383*.
- Tomsa 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. *arXiv preprint arXiv:2309.17332*.
- Souradip Guha, Ankan Mullick, Jatin Agrawal, Swetarekha Ram, Samir Ghui, Seung-Cheol Lee, Satadeep Bhattacharjee, and Pawan Goyal. 2021. Matscie: An automated tool for the generation of databases of methods and parameters used in the computational materials science literature. *Computational Materials Science (Comput. Mater. Sci.)*, 192:110325.
- Hiroaki Hayashi, Prashant Budania, Peng Wang, Chris Ackerson, Raj Neervannan, and Graham Neubig. 2021. Wikiasp: A dataset for multi-domain aspect-based summarization. *Transactions of the Association for Computational Linguistics*, 9:211–225.
- Fabian Kieser, Peter Wulff, Jochen Kuhn, and Stefan Küchemann. 2023. Educational data augmentation in physics education research using chatgpt. *Physical Review Physics Education Research*, 19(2):020150.

- 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. *arXiv preprint arXiv:1910.13461*.
- Zongxi Li, Xianming Li, Yuzhang Liu, Haoran Xie, Jing Li, Fu-lee Wang, Qing Li, and Xiaoqin Zhong. 2023. Label supervised llama finetuning. *arXiv preprint arXiv:2310.01208*.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2023. Calibrating llm-based evaluator. *arXiv preprint arXiv:2309.13308*.
- 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.
- Paula Maddigan and Teo Susnjak. 2023. Chat2vis: Generating data visualisations via natural language using chatgpt, codex and gpt-3 large language models. *IEEE Access*.
- Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems*, 35:462–477.
- Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.
- Rajdeep Mukherjee, Hari Chandana Peruri, Uppada Vishnu, Pawan Goyal, Sourangshu Bhattacharya, and Niloy Ganguly. 2020. Read what you need: Controllable aspect-based opinion summarization of tourist reviews. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 1825–1828.
- Ankan Mullick. 2023a. Exploring multilingual intent dynamics and applications. *IJCAI Doctoral Consortium*.
- Ankan Mullick. 2023b. Novel intent detection and active learning based classification (student abstract). *arXiv e-prints*, pages arXiv–2304.
- Ankan Mullick, Akash Ghosh, G Sai Chaitanya, Samir Ghui, Tapas Nayak, Seung-Cheol Lee, Satadeep Bhattacharjee, and Pawan Goyal. 2024. Matscire: Leveraging pointer networks to automate entity and relation extraction for material science knowledge-base construction. *Computational Materials Science*, 233:112659.
- Ankan Mullick, Surjodoy Ghosh D, Shivam Maheshwari, Srotaswini Sahoo, Suman Kalyan Maity, and Pawan Goyal. 2018a. Identifying opinion and fact subcategories from the social web. In *Proceedings of the 2018 ACM International Conference on Supporting Group Work*, pages 145–149.
- Ankan Mullick, Pawan Goyal, and Niloy Ganguly. 2016. A graphical framework to detect and categorize diverse opinions from online news. In *Proceedings of the Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 40–49.
- Ankan Mullick, Pawan Goyal, Niloy Ganguly, and Manish Gupta. 2017a. Extracting social lists from twitter. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 391–394.
- Ankan Mullick, Pawan Goyal, Niloy Ganguly, and Manish Gupta. 2018b. Harnessing twitter for answering opinion list queries. *IEEE Transactions on Computational Social Systems*, 5(4):1083–1095.
- Ankan Mullick, Shivam Maheshwari, Pawan Goyal, and Niloy Ganguly. 2017b. A generic opinion-fact classifier with application in understanding opinionatedness in various news section. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 827–828.
- Ankan Mullick, Ishani Mondal, Sourjyadip Ray, R Raghav, G Chaitanya, and Pawan Goyal. 2023. Intent identification and entity extraction for healthcare queries in indic languages. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1825–1836.
- Ankan Mullick, Abhilash Nandy, Manav Kapadnis, Sohan Patnaik, R Raghav, and Roshni Kar. 2022a. An evaluation framework for legal document summarization. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4747–4753.
- Ankan Mullick, Abhilash Nandy, Manav Nitin Kapadnis, Sohan Patnaik, and R Raghav. 2022b. Fine-grained intent classification in the legal domain. *arXiv preprint arXiv:2205.03509*.
- Ankan Mullick, Shubhraneel Pal, Tapas Nayak, Seung-Cheol Lee, Satadeep Bhattacharjee, and Pawan Goyal. 2022c. Using sentence-level classification helps entity extraction from material science literature. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4540–4545.
- Ankan Mullick, Sourav Pal, Projjal Chanda, Arijit Panigrahy, Anurag Bharadwaj, Siddhant Singh, and Tanmoy Dam. 2019. D-fj: Deep neural network based factuality judgment. *Technology*, 50:173.

- Ankan Mullick, Sukannya Purkayastha, Pawan Goyal, and Niloy Ganguly. 2022d. A framework to generate high-quality datapoints for multiple novel intent detection. *arXiv preprint arXiv:2205.02005*.
- 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.
- Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.
- Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.
- Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.
- Zhaoyi Sun, Hanley Ong, Patrick Kennedy, Liyan Tang, Shirley Chen, Jonathan Elias, Eugene Lucas, George Shih, and Yifan Peng. 2023. Evaluating gpt-4 on impressions generation in radiology reports. *Radiology*, 307(5):e231259.
- Jesse Vig, Alexander R Fabbri, Wojciech Kryściński, Chien-Sheng Wu, and Wenhao Liu. 2021. Exploring neural models for query-focused summarization. *arXiv preprint arXiv:2112.07637*.
- Yumo Xu and Mirella Lapata. 2020. Query focused multi-document summarization with distant supervision. *arXiv preprint arXiv:2004.03027*.
- Xianjun Yang, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Xiaoman Pan, Linda Petzold, and Dong Yu. 2022. Oasum: Large-scale open domain aspect-based summarization. *arXiv preprint arXiv:2212.09233*.
- Jiacheng Ye, Jiahui Gao, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. Progen: Progressive zero-shot dataset generation via in-context feedback. *arXiv preprint arXiv:2210.12329*.
- Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: A tale of diversity and bias. *arXiv preprint arXiv:2306.15895*.
- Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
- Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don’t make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*.
- Haichao Zhu, Li Dong, Furu Wei, Bing Qin, and Ting Liu. 2022. Transforming wikipedia into augmented data for query-focused summarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2357–2367.

Appendix

A Related Work

In this section, we conduct a survey of state-of-the-art literature on closely related topics.

Summarization: We perform a survey of various summarization techniques, that encompasses generic approaches such as abstractive (Chopra et al., 2016; See et al., 2017; Paulus et al., 2017) and extractive summarization like TextRank (Mihalcea and Tarau, 2004) and LexRank (Erkan and Radev, 2004). Subsequent research has explored summarization in specific directions, including aspect-based summaries (Hayashi et al., 2021; Coavoux et al., 2019; Mukherjee et al., 2020; Akhtar et al., 2017) and query-focused summaries (Vig et al., 2021; Zhu et al., 2022; Xu and Lapata, 2020), but none focuses on domain-specific persona-based summarization.

Persona Concept: (Goldsack et al., 2023; Luo et al., 2022) focus on building lay-summarization for comprehensible to non-technical audiences but do not have the differentiating factor of persona concept to distinguish the various technical summaries. based on persona and no usage of LLMs as an alternative metrics to evaluate the summaries. Our work differs from the fact that we show a pipelined approach - data generation, persona-specific (doctor vs patient vs normal person) summarization with key points and GPT-4 based evaluation to save time and assist different persona to augment their knowledge on it to make conclusions more efficiently. The concept of Persona also different from the notion of intent (Mullick et al., 2022d, 2023, 2022b; Mullick, 2023b,a; Mullick et al., 2022a), entity-relation (Mullick et al., 2024; Guha et al., 2021; Mullick et al., 2022c) or opinion/fact (Mullick et al., 2017b, 2016, 2018a,b, 2019, 2017a) idea. Our work differs from the fact that we show a pipelined approach - data generation, ‘persona concept’ specific (doctor vs patient vs normal person) summarization with key points and GPT-4 based evaluation to save time and assist different persona to augment their knowledge on it to make conclusions more efficiently.

Data Generation using LLMs: Large Language Models (LLMs), such as the GPT family, have been utilized to generate training datasets for NLP tasks, addressing data scarcity issues in a cost effective manner (Yu et al., 2023; Meng et al., 2022;

Ye et al., 2022). ChatGPT¹⁰ aids in educational data augmentation and data visualization (Kieser et al., 2023; Maddigan and Susnjak, 2023); however, in healthcare domain, GPT-3 and GPT-4 are employed for generating medical dialogue summarization data (Chintagunta et al., 2021) and radiology reports (Sun et al., 2023). However, no prior work focuses on benchmark domain-specific persona-based summary data generation with appropriate human validation.

LLMs as Evaluation Metrics: The rise of LLMs presents a potential, cost-effective alternative for evaluating various NLP tasks. Existing efforts include a taxonomy of LLM-based NLG evaluation methods (Gao et al., 2024), the development of ‘ChatEval’ for assessing response quality (Chan et al., 2023), and proposed guidelines for LLM-based evaluation (Zhou et al., 2023). While some studies explore LLM-based assessments with human alignment (Liu et al., 2023), there is a lack of work utilizing LLMs to evaluate solution architectures comprehensively with a critique scoring system.

Our paper takes a significant step towards addressing the shortcomings of prior literature in terms of persona-based summary data generation followed by human validation as well as performing LLM based critic evaluation.

B Dataset Requirement

Access to robust, comprehensive datasets is crucial for training NLP models to understand and generate persona-specific content, emphasizing the challenges of data scarcity and summary-making capabilities. Manual annotation of large healthcare datasets is both costly and time-intensive, demanding domain expertise and meticulous attention to detail. Consider the example where n individuals generate summaries for m persona, resulting in $n \times m$ distinct summary generations for a single article — it is a highly expensive and time consuming process. Moreover, acquiring suitable human resources for labeling is challenging, as people are often hesitant to undertake the tedious and difficult task of summary generation, even with a standard payment agreement.

We provide healthcare data to the Prolific annotation platform with specific criteria: annotators

¹⁰<https://chat.openai.com/>

with a PhD / Graduate degree, a medical background, approval rates of 90%-100%, and expertise in medicine. However, out of 157,341 potential annotators, only 189 meet these criteria, and even among them, there is a high rejection rate of 71.43% for selecting documents for manual summary generation, despite offering more than standard payment. Further, eligible annotators tend to heavily rely on ChatGPT and similar automated approaches, leading to the need for extensive re-evaluations and revisions, along with issues related to annotator rejection, even to assess a limited number of documents. Hence, obtaining a high quality dataset remains a formidable challenge.

C GPT-4 Bias

It is acknowledged that using summaries generated exclusively by GPT-4 could introduce biases inherent in its summarization capabilities, it may also be noted that alternatives, such as human evaluation, also carry their own biases. Despite the potential for bias, leveraging GPT-4 for summarization may still be a pragmatic choice, especially in scenarios where access to diverse datasets or sophisticated validation methods is limited. However, in this work, we remain vigilant, recognizing the limitations inherent in both automated and human-generated summaries, and take proactive steps such as human intervention to validate and contextualise the results to mitigate biases to the best extent possible within the given constraints.

D Prompts

We use prompting in three stages - data generation, finetune-inference and critique. There are two kinds of prompts - system prompt and user prompt.

D.1 Data Generation

system : You are an AI assistant who are to generate a summary of a medical document specific to a certain persona which can be doctor, patient, normal person. The summary of a medical document should be generated from the perspective of the respective persona.

user : Summarize the medical document given below from the perspective of a {persona} [doctor/patient/normal person] and return the summary only. The medical document is as follows: Document: {document}

D.2 Finetune and Inference prompt

user prompt - Summarize the medical document given below from the perspective of a persona:

Document: document

D.3 Critique

system: You are an AI assistant who is to evaluate the summary of a medical document specific to a certain persona which can be doctor, patient or a normal person. A doctor requires a detailed and technical summary about the medical document. Patients require a layman's summary about the medical document, with information about things like causes, effects, treatment etc. that may be helpful to them. A normal person has no medical knowledge and requires a generic summary about the medical document. You need to return a score between 0 and 1 reflecting the quality of the generated summary based on some criteria.

user: You are given a medical document and the corresponding summary of the document generated from the perspective of a {persona} predicted by a language model as follows.

Document: {document}

Ground truth summary : {label summary}

Summary from the perspective of a {persona} [doctor/patient/normal person]: {model generated summary}

Evaluate the above persona based summary for the document in terms of each of the following criteria and return only a score between 0 and 1 without any explanation:

- The extent to which the generated summary is relevant to the specific persona {persona}[doctor/patient/normal person] based summary of the document.
- The extent to which the generated persona-based summary correctly covers all the important key points described in the persona {persona}[doctor/patient/normal person] based summary of the document.
- The extent to which the summary does not contain information specific to all other possible personas {persona_set - persona}[doctor/patient/normal person] based summary.
- Rate the summary from the point of view of the persona – whether the summary is good, average, or bad. A good summary effectively

captures the essential points, presenting them clearly and concisely. It maintains accuracy, encourages reader engagement, and serves as a compelling introduction to the content. An average summary conveys the main points but may lack some clarity or detail, presenting a decent overview without standing out in terms of conciseness or precision. It provides a basic understanding but not from a more refined focused summary and fails to accurately convey the main points, containing inaccuracies or misinterpretations. It is either overly verbose or lacks coherence, making it difficult for the reader to grasp the core information effectively.

- Calculated summary from the point of view of the persona [Good/Bad/Average] [Calculated from 4 with the help of manual annotation]

E Experiments

E.1 Varying Training Size Dataset

To understand the effect of training data size on the performance, we vary the WebMD training data for the Llama2-13b model - taking 10-shot (k -shot settings where $k=10$), 10%, 40% and 70% of the initial training data, and finetune the Llama2-13b model with same parameter and hyper-parameter settings and the five criterias of GPT-4 Critique outcome (in %) are shown in Fig 1. We see that with increasing the dataset size, the performance of Llama2-13b improves in terms of GPT-4 critique and traditional metrics. Even at 40% of the dataset, the model is able to achieve a very good performance. It shows the effectiveness of the Llama2-13b model. It also infers that even with very little amount of data in 10-shot Llama2-13b can able to generate appropriate persona and aspect based summary.

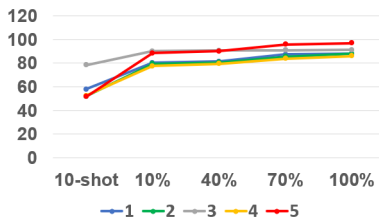


Figure 1: Different Training Data Sizes

E.2 Tuning generation parameters during model inference:

We investigate the impact of tuning the *max-new-token* and *temperature* generation parameters on

the performance of our finetuned model during inference. The variation in performance in terms of the five GPT-4 critique based criteria are shown in Figures 2a and 2b respectively. We observe that our model performs the best for a temperature of 0 and performance degrades significantly as we increase the temperature beyond 0.4. Similarly, the best model performance is achieved for a *max-new-token* size of 350. We have used NLTK, Spacy, openai (version=0.28), huggingface_hub, torch and transformers python packages for all experiment.

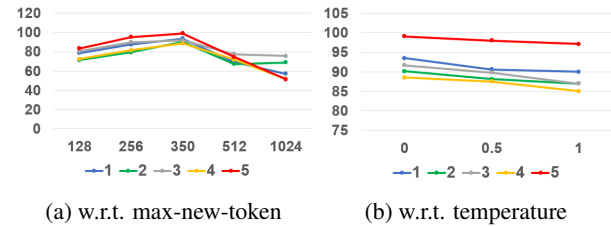


Figure 2: Variations of Llama2-13b-Finetune (in %)

E.3 Time and GPU

We experiment on 80GB A100 GPU with GPU clock cycle 210 MHz. The finetuning and inference time of our finetuned models are in Table 6.

| Model | Finetune Time | Inference Time |
|------------|---------------|----------------|
| Llama2-7b | 8 hrs | 2hrs 30 mins |
| Llama2-13b | 20 hrs | 2hrs 50 mins |

Table 6: Model Training Time [using 80GB A100 GPU]

F Human Annotations

Annotation Guidelines for Comparative Rating:

We provide instructions with explanations of different *persona* and ask to identify which summary belongs to which persona as shown in Fig 5. We also provide the link of document along with distinct summaries of GPT-4 and Llama2-13b finetune model for comparison as shown in Fig 6. The instructions are the following -

“You are given a summary of a medical document specific to the perspective of a certain group of people (doctor, patient and normal person).

A doctor requires a detailed and technical summary about the medical document.

A patient requires a layman summary about the medical document, with information about things like causes, effects, treatment etc. that may be helpful to him. A patient only requires a top level view of the extensive medical details and not so much

medical details like a doctor.

A normal person has no medical knowledge and requires a generic summary about the medical document and does not require extensive medical details.”

Annotation Guidelines to Prolific and Doctors

For annotator selection, we have several criterias. Annotator selection includes specific criteria such as ‘Degree subject’ in Health and welfare, ‘Highest education level completed’ as Doctorate degree or Graduate degree, ‘Fluent languages’ in English, ‘Approval Rate’ of 90–100, ‘Subject’ in Medicine, and ‘Employment Sector’ as Doctor. Further annotations are conducted by graduate doctors (details in Section - Human Evaluation Part).

Following is the annotation guideline to ‘Prolific’ annotation platform -

Objective: Generate Personified Summaries by Prolific

Introduction: In this study, you are tasked with generating a summary tailored to three different personas: a Doctor, a Patient, and a Normal Person. You will be provided with a document link containing a Source Document (SD) – which can be a medical research document or a general article related to health from <https://www.webmd.com/>. . Additionally, a Persona (P) will be present.

Your Task: Read the Source Document (and Persona) and craft three summaries, each targeted towards one of the personas mentioned. Use your understanding and perspective to tailor the information in a way that is most relevant and comprehensible to each persona.

Summary Persona:

Doctor Persona: Craft a summary that focuses on medical terminology, guidelines, and provides information suitable for a medical professional. Emphasize technical accuracy and relevance to medical practice. A doctor requires a detailed and technical summary about the medical document.

Patient Persona: Generate a summary with a patient-centric approach, avoiding excessive technical jargon. Ensure that the information is clear, easily understandable, and addresses concerns that a patient might have. A patient requires a non-technical summary about the medical document, with information about things like causes, effects, treatment etc. that may be helpful to him. A patient only requires a top level view of the extensive medical details and not so much medical details like a doctor.

Normal Person Persona: Tailor a summary for

a general audience without a medical background. Use simple language, avoid technical terms, and present the information in a way that is accessible and engaging to a layperson. A normal person has no medical knowledge and requires a generic summary about the medical document and does not require extensive medical details.

Instructions: 1. Carefully review the Source Document and the Persona. Consider the specific needs and understanding level of each persona while generating the summaries.

2. No additional software download is required. Use a browser, preferably Google Chrome, and ensure a stable internet connection.

3. Allocate time judiciously for crafting each of the three summaries based on the provided 2 SD instances.

4. After completion, you will be asked to provide feedback on the generation exercise, platform interaction, and details about your academic background, age, country of birth, and any medical background or experience with model-generated summaries.

Payment Requirements: Upon completing the study, click on the provided link containing the completion code to redirect you to the Prolific platform. Payment will be processed within one to two weeks.

Ethical Considerations:

Adhere to strict confidentiality and data protection standards to ensure the privacy of medical information. If you have concerns or questions, feel free to reach out, as this study aligns with ethical guidelines.

This study aims to harness diverse perspectives, including those of medical professionals, to refine the generation of personified summaries for enhanced utility in various contexts.

Next, the details while providing the documents - “You will be given 2 documents in the next 2 pages and you need to write the summaries with respect to Doctor, Patient and Normal Person (as in example).

Please do not use ChatGPT/GPT-4 or any Large Language Models - all summaries should be generated by human properly. It is a strict instruction and will be checked manually - if found any issue: it will be rejected and re-doing will be required.

Your summary length (word count) is approximately 15% - 20% of the document length (word count) for three different types.”

G Examples

Two human annotated examples in doctor persona is shown in Fig 3 where the GPT-4 generated summary is better and Fig 4 where LLAMA-2 generated summary is better. Two examples of human annotation interface is shown in Fig 5 and Fig 6 respectively.

Doc link - <https://www.webmd.com/cholesterol-management/what-is-cholesterol-test>

GPT- 4 Summary- The medical document discusses the importance, procedures, and interpretations of cholesterol tests. These tests are crucial in evaluating the risk of heart attack or stroke by measuring cholesterol and triglyceride levels in the blood. The test is recommended every five years from age 20, with increased frequency for individuals over 45 (men) or 50 (women), those with cholesterol greater than 200, low HDL cholesterol, obesity, high blood pressure, other high-risk conditions, or those undergoing treatment for high cholesterol. The American Heart Association recommends conducting the test in a doctor's office, hospital, or lab to ensure accuracy and personalized follow-up care. Results interpretation includes total cholesterol (healthy level: 200mg/dL or less), LDL bad cholesterol (healthy level: less than 100mg/dL), HDL good cholesterol (healthy level: 60mg/dL or higher for both genders), and triglyceride levels (healthy level: less than 150mg/dL). These numbers are not universally applicable; the doctor will define specific targets based on the patient's health condition. If the results are in the unhealthy range, lifestyle modifications such as diet and exercise changes are suggested initially, followed by medication if necessary. If all results are within healthy limits, the test is usually not repeated for another five years.

LLAMA-2 Summary- The document discusses the importance of regular cholesterol testing for individuals aged 20 and above. It emphasizes the need for testing every five years or more frequently for those with high cholesterol levels, low HDL levels, obesity, high blood pressure, or other conditions that increase the risk of heart disease. The document also mentions that the test should be conducted in a medical setting, either at the doctor's office, hospital, or laboratory, to ensure accurate results. The test results provide a lipid panel, which includes the total cholesterol level, LDL and HDL levels, and triglyceride levels. A low LDL level is considered healthy and protects against heart attacks and strokes. A high HDL level is considered healthy and protects against heart disease. A low triglyceride level is also considered healthy. The document also mentions that the target levels for each of these parameters may vary depending on the individual's health status. If any of the test results fall in the unhealthy range, the individual may be advised to make lifestyle changes or take medication to improve their health. The document emphasizes the importance of maintaining a healthy lifestyle, including a balanced diet and regular physical activity, to prevent heart disease.

Figure 3: GPT-4 generated summary better than LLAMA2-13b model generated summary[persona : doctor]

Doc link- <https://www.webmd.com/sex/birth-control/nausea-from-birth-control-pills>

GPT-4 Summary- The document addresses the side effects, specifically nausea, related to birth control pills and offers strategies to mitigate these symptoms. Birth control pills are an effective means of pregnancy prevention, with additional benefits of lighter periods and improved acne, possible prevention of certain types of breast disease, anemia, ovarian cysts, and ovarian and endometrial cancers. However, nausea is a common drawback due to the estrogen in the pill irritating the stomach lining, potentially leading to vomiting, bloating, and loss of appetite. If vomiting occurs within 2 hours of pill intake, contraceptive absorption may not have occurred and another pill should be taken. The document suggests several methods to prevent pill-induced nausea including taking the pill at a consistent time, preferably at night, consuming light, bland foods, drinking cold liquids, eating smaller, slow meals, avoiding post-eating activities and spicy foods. If these methods fail, consultation with a healthcare provider about alternative contraceptive methods is advised.

LLAMA-2 Summary- The medical document highlights the benefits and drawbacks of birth control pills, including their effectiveness in preventing pregnancy, regulating periods, improving skin conditions, and potentially reducing the risk of certain cancers. However, it also mentions that nausea can be a common side effect of these pills, especially during the first three months of use. The document suggests that the hormones in the birth control pills, particularly estrogen, can irritate the lining of the stomach, leading to symptoms such as nausea, vomiting, bloating, and loss of appetite. If a pill is taken within two hours of vomiting, it may not be effective in preventing pregnancy. The document recommends several strategies to help prevent or manage nausea, including taking the pill at the same time every day, eating light, bland foods, drinking cold liquids, eating smaller meals slowly, and avoiding spicy foods. If nausea persists, it suggests consulting a doctor about switching to a different birth control pill or using alternative contraceptive methods. Overall, the document emphasizes the importance of weighing the benefits and drawbacks of birth control pills and finding the right method that works best for an individual's needs and lifestyle. doctor : The patient is experiencing nausea due to the birth control pills. The patient can take antacid tablets to manage the nausea. However, if the nausea persists, it is recommended to consult a doctor.

Figure 4: LLAMA2-13b model generated summary better than GPT-4 generated summary[persona : doctor]

Document link - <https://www.webmd.com/ibd-crohns-disease/ulcerative-colitis/antineutrophil-cytoplasmic-antibodies-anca-test>

Summary -

An antineutrophil cytoplasmic antibodies (ANCA) test is a blood test used to check for proteins your immune system makes to fight germs. This test can help your doctor distinguish between ulcerative colitis and Crohn's disease, and diagnose autoimmune diseases like vasculitis. In some cases, your immune system may mistakenly attack your healthy tissues, these are called autoantibodies. The presence of these autoantibodies in your blood could indicate you have an autoimmune disease. The ANCA test checks for these autoantibodies. This test is particularly useful if your doctor is unsure whether you have Crohn's disease or ulcerative colitis, as the symptoms of these two types of inflammatory bowel disease are often similar. The ANCA test can help confirm a diagnosis when other tests are inconclusive. Up to 80% of people with ulcerative colitis test positive for a specific type of ANCA. The ANCA test can also predict whether you'll respond to certain medications for inflammatory bowel disease and is commonly used to diagnose vasculitis, a group of autoimmune disorders that affect small blood vessels. In vasculitis, ANCA can cause an attack on the blood vessels, causing them to swell. Depending on which blood vessels are affected, vasculitis can cause a variety of symptoms, including fever, tiredness, weight loss, muscle or joint aches, vision or hearing loss, skin rashes, and changes in urine. The ANCA test can also show if your treatment for vasculitis is working and can predict if symptoms might come back after treatment. The test is done by drawing a blood sample, which is then sent to a lab for analysis. If the test is positive, you may have an autoimmune vasculitis or, if you are having symptoms of inflammatory bowel disease, you are more likely to have ulcerative colitis rather than Crohn's disease. Depending on the results, your doctor will discuss the next steps, including potential further tests or treatments.

- Doctor
- Patient
- Normal Person

Figure 5: Persona identify experiment example snapshot

Document link - <https://www.webmd.com/sex/birth-control/nausea-from-birth-control-pills>

Patient's Summary 1-

Birth control pills are a safe and effective method for preventing pregnancy. They are also known to lighten periods, improve acne, and prevent certain diseases. However, one potential side effect is nausea, which is often caused by the hormone estrogen irritating the stomach lining. If you vomit within 2 hours of taking the pill, it may not have been absorbed properly and you may need to take another one. To prevent nausea from birth control pills, try taking the pill with a meal or using antacids. Other helpful tips include taking the pill at the same time each day, preferably at night before bed, eating light and bland foods, drinking cold liquids, eating smaller meals slowly, and avoiding activity and spicy foods after eating. If these strategies don't help, consult your doctor about potentially switching to a different pill or another type of contraceptive.

Patient's Summary 2-

The document explains that birth control pills are a safe and effective way to prevent pregnancy, but they can also cause nausea. Nausea is usually caused by the hormones in the pill, which can irritate the lining of the stomach. The document suggests several ways to prevent and treat nausea, such as taking the pill at the same time every day, eating light, bland foods, drinking cold liquids, eating smaller meals slowly, avoiding activity after eating, and avoiding spicy foods. If these methods don't work, it's suggested to talk to a doctor about switching to a different birth control pill or using other forms of contraception. patient : Birth control pills are a safe and effective way to prevent pregnancy. They can make periods lighter and improve acne, and may help prevent certain types of breast disease, anemia, ovarian cysts, and ovarian and endometrial cancers. However, they can also cause nausea. This is usually caused by the hormones in the pill, which can irritate the lining of the stomach. Nausea can be treated with antacid tablets, but it can also be prevented with a few lifestyle and diet changes. These include taking the pill at the same time every day, eating light, bland foods, drinking cold liquids, eating smaller meals slowly, avoiding activity after eating, and avoiding spicy foods.

- Both are good
- Summary 1 is better
- Summary 2 is better
- Both are bad

Figure 6: Llama2-13b finetune and GPT-4 summary comparison experiment example snapshot