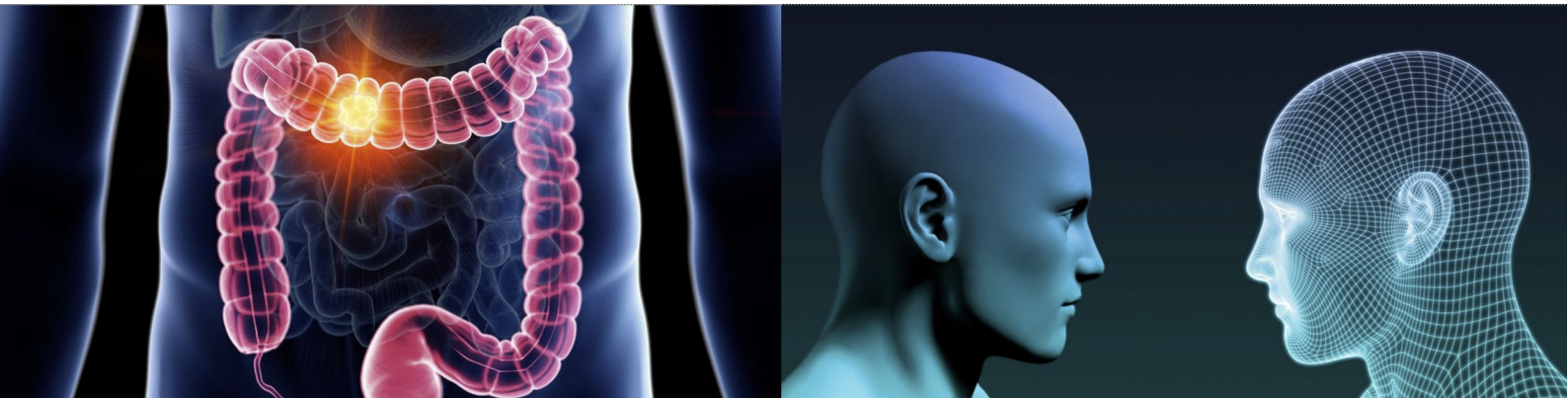


Exploring Hybrid Intelligence for Topic Interpretation in Colorectal Cancer Research: A Comparative Study of GPT-3.5 and Human Expertise



Ayush Patandin

Exploring Hybrid Intelligence for Topic Interpretation in Colorectal Cancer Research: A Comparative Study of GPT-3.5 and Human Expertise

THESIS

submitted in fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

by

Ayush Patandin
born in Paramaribo, Suriname

To be defended publicly on Wednesday August 30, 2023 at 15:30 PM (CET)



Web Information Systems Research Group
Department of Software Technology
Faculty EEMCS, Delft University of Technology
Delft, the Netherlands
<https://www.wis.ewi.tudelft.nl/>



Erasmus MC
Dr. Molewaterplein 40, 3015 GD
Rotterdam, the Netherlands
<https://www.erasmusmc.nl/en/>

© 2023 Ayush Patandin.

Cover pictures: <https://skywell.software/wp-content/uploads/2019/10/ai-vs-human-1024x512.jpg>

<https://surgery.wustl.edu/wp-content/uploads/2021/06/News-Colon-Cancer-Screening-1.png>

Exploring Hybrid Intelligence for Topic Interpretation in Colorectal Cancer Research: A Comparative Study of GPT-3.5 and Human Expertise

Author: Ayush Patandin
Student id: 4958195
Email: S.A.S.patandin@student.tudelft.nl

Abstract

Colorectal cancer is a widespread disease that significantly impacts the health of individuals worldwide. Understanding the needs and concerns of those affected by this disease is crucial for improving patient outcomes and enhancing the quality of care. Patient web forums have emerged as valuable platforms for individuals to openly share their experiences and thoughts related to colorectal cancer, providing unique insights into the social, physical and emotional aspects of their patient journey. These forums offer a more comprehensive and authentic portrayal of patient experiences compared to traditional patient data collection methods, such as questionnaires and interviews, which may not capture the full scope of patients experiences in the colorectal cancer carepath.

However, analyzing the vast amount of unstructured data within these patient web forums presents a significant challenge. Traditional manual analysis by human experts is time-consuming, labor-intensive, and limited in scalability, making it impractical to analyze the sheer volume of patient-generated content. This is where the application of natural language processing (NLP) techniques becomes crucial. NLP enables the automated processing and analysis of textual data, allowing for efficient extraction and interpretation of the large amounts of patient forum posts.

Nevertheless, relying solely on machine intelligence, such as topic modeling and natural language generation, for interpreting patient forum data carries inherent risks, including the potential for disseminating misleading information. While these machine-driven techniques offer efficient and scalable ways to analyze and generate insights from the large amount of diverse and unstructured patient forums, they may lack the necessary contextual understanding and domain expertise to ensure the accuracy, relevance, and ethical implications for interpreting colorectal cancer patient experiences.

To close this gap between human experts and machine intelligence, this thesis explores the potential of hybrid intelligence (HI) for topic interpretation in colorectal cancer research. The main research question is: “How can topic modeling, GPT-3.5 language generation and human expertise be combined to explore the interpretation of patient web forums in colorectal cancer (CRC) research?”

To address the research question, three human studies were conducted. The first study employed NMF topic modeling to compare topic interpretations created independently by medical workers and GPT-3.5. This comparative analysis discovered unique observations that differentiate human-written and AI-generated interpretations on online patient stories. In the second study, it was investigated how medical researchers collaborate with GPT-3.5 to develop hybrid interpretations on patient experience topics generated by the BERTopic model. A Flask web application served as the interactive platform for combining their knowledge with the AI model. Finally, the third study made professional human evaluators assess the topic relevance of the interpretations generated by medical researchers and GPT-3.5 to determine whether the combination of GPT-3.5 and human expertise leads to improved topic interpretations compared to individual interpretations.

The proposed solution to the research problem is to explore a hybrid workflow that compares, combines and validates GPT-3.5 language generation and human expertise, aiming for enhanced interpretations of topics extracted from colorectal cancer patient forums. The three studies provide opportunities for researchers and medical professionals to integrate machine intelligence from topic models and GPT-3.5 in their field of work. The hybrid workflow has conclusively demonstrated that human experts were successfully able to compare and enhance the relevance of human and GPT-3.5 interpretations of colorectal cancer patient experience topics. This allowed human experts to efficiently reach a more comprehensive understanding of patient forum data, which is essential for improving patient health in colorectal cancer research.

Thesis Committee:

Chair:	Dr. ir. Christoph Lofi, Web Information Systems, TU Delft
University supervisor:	Dr. ir. Jie Yang, Web Information Systems, TU Delft
Company supervisor:	Dr. ir. Jiwon Jung, Erasmus MC
Committee Member:	Dr. ir. Johan Pouwelse, Distributed Systems, TU Delft

Preface

In this report, I present you with my graduation thesis titled “Exploring Hybrid Intelligence for Topic Interpretation in Colorectal Cancer Research: A Comparative Study of GPT-3.5 and Human Expertise”. This work is done to fulfill the requirements for the degree of Master of Science in Computer Science at Delft University of Technology (TU Delft). The main contributions of this project were held at Erasmus MC, Rotterdam. Throughout the period from January 16, 2023, to August 30, 2023, I performed my scientific work under the guidance of Dr. ir. Jie Yang (TU Delft) and Dr. ir. Jiwon Jung (TU Delft, Erasmus MC).

During this trajectory of 8 to 9 months, I was given the opportunity to integrate elements from my specialization track Data Science and Technology in the healthcare domain. I also made myself comfortable to work in the same environment as medical doctors at the surgery department of Erasmus MC. Throughout my thesis journey, I had the privilege of collaborating with Yingtao Sun, a fellow Master’s student from the Faculty of Industrial Design at TU Delft. Both of our research work centered around the colorectal cancer disease, and therefore, it was really beneficial to hold discussions on the same topics.

Of course, I am grateful for having done my thesis under the supervision of Dr. ir. Jie Yang and Dr. ir. Jiwon Jung. Thanks to their guidance during our multiple meetings, I was able to finetune my research directions and application work to achieve my goals for this thesis. I also want to thank Dr. ir. Christoph Lofi and Dr. ir. Johan Pouwelse for being part of my thesis committee. In addition, I want to thank all of the medical researchers and healthcare providers from Erasmus MC who took part in the human studies of my project. Due to their involvement, it became possible to explore their domain-specific knowledge together with my data-driven analysis.

Lastly, I cannot go without expressing my deepest appreciation to my parents, siblings, and friends. I am forever grateful for their motivation and support, especially during difficult times. After this beautiful chapter of my life, I am determined to make the most of every opportunity that comes my way and continue to make them proud.

Ayush Patandin
Delft, the Netherlands
August 22, 2023

Contents

Preface	iii
Contents	v
List of Figures	vii
List of Tables	ix
List of Abbreviations	xi
1 Introduction	1
1.1 Research Questions	3
1.2 Thesis Objectives	3
1.3 Thesis Outline	4
2 Related Work	5
2.1 Topic Modeling	5
2.2 Interpretation with Pre-trained Language Models	7
2.3 Towards Hybrid Intelligence	10
3 Methodologies	13
3.1 Data Collection	13
3.2 Data Analysis	15
3.3 Topic Interpretation with GPT-3.5	22
3.4 Design-based Topic Interpretation	24
4 Comparative Analysis of Human and AI Interpretations	27
4.1 Study Procedure	27
4.2 Study Reflection	29
5 Developing Interpretive Agreements	35
5.1 Study Procedure	35

CONTENTS

5.2	Study Reflection	39
6	Expert Evaluation of Hybrid and Individual Interpretations	47
6.1	Study Procedure	47
6.2	Study Reflection	50
7	Conclusions and Future Work	53
7.1	Discussions	53
7.2	Conclusion	55
7.3	Future work	56
	Bibliography	59
A	Patient Journey Map	67
B	First Study’s Individual Human and AI NMF Topic Interpretations	69
C	Second Study’s Flask Webpages	74
D	Second Study’s Saved Interpretations per Agreement Category	77
E	Third Study’s Interpretation Evaluation Results	81

List of Figures

1.1	Three-dimensional theoretical framework of health [45, Figure 1].	1
2.1	Caption for LOF	9
2.2	Demonstration of the positive and negative aspects of human cognition and artificial intelligence [19, Figure 2.1].	11
2.3	Distribution of roles in hybrid intelligence [24, Figure 2].	12
3.1	Brief overview of CRC discussions on Cancer Survivors Network USA.	14
3.2	Ideal number of LDA topics based on Maximum Difference between Topic Coherence and Topic Overlap (Mean Jaccard Similarity).	17
3.3	BERTopic baseline constructed with 100 topics and the sub-models mentioned on the right side of the arrow.	19
3.4	Caption for LOF	20
3.5	Similarity Heatmap between the 50 NMF topics used in the first study and the 100 BERTopic topics employed in the second and third study.	21
3.6	Examples of Topic Interpretation Generation by the 'text-davinci-003' GPT-3.5 Model Release. The significance of topic terms in the word cloud is reflected by their respective sizes, with larger terms indicating higher relevance to the topic.	23
3.7	The NMF generated upper clusters along with their respective sizes indicating how many documents are contained within each cluster.	24
4.1	Summary of Participant Work Experience.	28
4.2	Example of a topic card and a comment card used for human expert interpretations on NMF topics during the co-creation session of study 1.	29
4.3	Summary of evaluated topics per participant.	31
4.4	Example observation of GPT-3.5's behavioral change for interpreting diverse topic information.	33
5.1	Flow diagram for combining GPT-3.5 and doctor interpretations through Flask Web Application.	36
5.2	Zero-shot Question Answering task to get Topic Information from GPT-3.5.	37

LIST OF FIGURES

5.3	Zero-shot text-to-text transformation task to refine GPT-3.5 topic description based on Expert Feedback.	38
5.4	Summary of Participant Expertise and Flask User Experience.	40
5.5	Distribution of Methods for Reaching Interpretive Agreements.	42
5.6	Human-AI Similarity and Response Time Measurements for each type of Interpretive Agreement.	43
5.7	Distributions of refinement methods by human experts on human and AI interpretations.	44
5.8	Summary of Human and AI refinements for different agreement categories. . .	45
6.1	Survey construction for 3 human evaluators using the interpretations from study 2.	48
6.2	Example of participant task for the given set of interpretations.	49
7.1	Hybrid workflow overview: step-by-step analysis for comparing, combining and validating Human and AI Interpretations in understanding CRC patient experiences.	56
A.1	Designer’s Patient Journey Map after she interpreted the NMF topic modeling results for the first study	67
B.1	Participant 1 - Cluster 4 Interpretations	69
B.2	Participant 1 - Cluster 8 Interpretations	69
B.3	Participant 2 - Cluster 7 Interpretations	70
B.4	Participant 2 - Cluster 10 Interpretations	70
B.5	Participant 3 - Cluster 4 Interpretations	70
B.6	Participant 4 - Cluster 4 Interpretations	71
B.7	Participant 4 - Cluster 10 Interpretations	71
B.8	Participant 5 - Cluster 5 Interpretations	71
B.9	Participant 5 - Cluster 8 Interpretations	72
C.1	Flask Web Application used to develop hybrid interpretations as part of the second study	74
C.1	Flask Web Application used to develop hybrid interpretations as part of the second study (cont.)	75
D.1	Saved Interpretations for ‘same_both’ (No refinements)	77
D.2	Saved Interpretations for ‘same_gpt’ (Only refinements made by Human) . . .	78
D.3	Saved Interpretations for ‘same_human’ (Only refinements made by GPT-3.5) .	79
D.4	Saved Interpretations for ‘different_both’ (Refinements made by both human and GPT-3.5)	80

List of Tables

3.1	Cancer Survivors Network USA CRC dataset attributes	15
3.2	Comparison between NMF and LDA based on C_{NPMI} coherence score and Average Jaccard Similarity of the top 20 topic terms for different numbers of topics.	18
3.3	BERTopic c_V coherence scores for different numbers of topics using the same model configurations of Figure 3.3.	19
3.4	Summary of tasks performed by the designer for the NMF and BERTopic baselines.	25
4.1	Selected clusters for each participant.	30
4.2	Scenario-based observations that differentiate how human experts and GPT-3.5 interpret NMF topic modeling results.	34
5.1	Saved attributes for each topic assessed by the participants.	39
5.2	Summary of Numerical Analysis of Interpretive Agreement Methods.	43
6.1	Average percentage ratings of each participant survey for original human, original GPT-3.5, improved human and improved GPT-3.5 topic interpretations	50
6.2	Global average percentage ratings for original human, original GPT-3.5, improved human and improved GPT-3.5 topic interpretations	51
E.1	Expert Evaluation Results: Interpretation ratings on a scale of 1 to 7 from professional Human Evaluators at Erasmus MC	82

List of Abbreviations

AI	Artificial Intelligence.
API	Application Programming Interface.
BERT	Bidirectional Encoder Representations from Transformers.
BLEU	BiLingual Evaluation Understudy.
CRC	Colorectal Cancer.
GPT	Generative Pre-trained Transformer.
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise.
HI	Hybrid Intelligence.
HTML	Hypertext Markup Language.
JSON	JavaScript Object Notation.
LDA	Latent Dirichlet Allocation.
LLM	Large Language Model.
MC	Medical Center.
MMR	Maximal Marginal Relevance.
NER	Named Entity Recognition.
NLG	Natural Language Generation.
NLP	Natural Language Processing.
NLTK	Natural Language Toolkit.
NLU	Natural Language Understanding.
NMF	Non-negative Matrix Factorization.
NPMI	Normalized Pointwise Mutual Information.
PhD	Doctor of Philosophy.
PII	Personal Identifiable Information.
PLM	Pre-trained Language Model.
Regex	Regular Expression.
ROUGE	Recall-Oriented Understudy for Gisting Evaluation.
STS	Semantic Textual Similarity.
TF-IDF	Term Frequency-Inverse Document Frequency.
UMAP	Uniform Manifold Approximation and Projection.
URL	Uniform Resource Locator.
USA	United States of America.

Chapter 1

Introduction

As one of the most widespread forms of cancer globally, colorectal cancer (CRC) can have a significant impact on patients' lives [13]. Early diagnosis and proper treatment are crucial factors that can improve patient outcomes and enhance the quality of care. To achieve this, healthcare providers must gain knowledge by understanding the experiences of CRC patients, including their side-effects, activities, and support from family and friends. This makes it more important than ever for medical workers to concentrate on patient centeredness as a quality domain for improving healthcare [42].

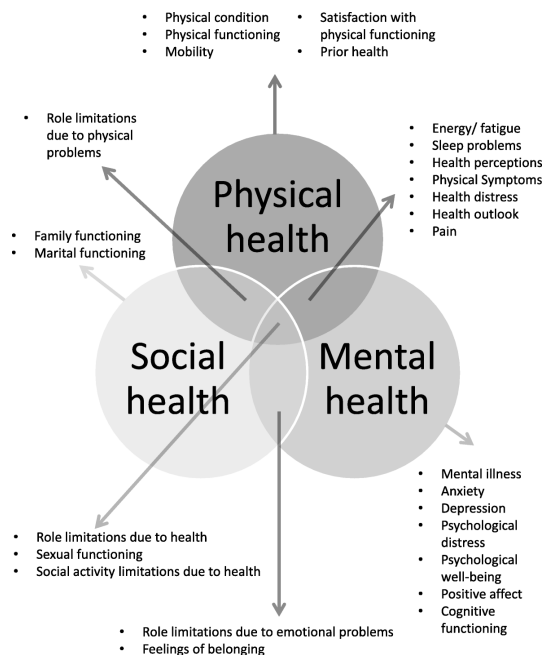


Figure 1.1: Three-dimensional theoretical framework of health [45, Figure 1].

Researchers have long strived to gain a comprehensive understanding of the diverse experiences and challenges faced by CRC patients in managing their health. To facilitate this understanding, they have categorized patient health into physical, mental, and social

aspects, as illustrated in Figure 1.1 [45, Figure 1]. However, despite these efforts, truly grasping and interpreting the full spectrum of patient experiences remains a difficult task, particularly when relying on traditional data collection methods. While controlled clinical trials provide valuable insights, they may not fully capture the wide range and depth of patient experiences [27]. Patients may feel hesitant to discuss the various impacts of their treatment or may be constrained by the limited scope of structured questionnaires and interviews. For instance, extensive surveys may prompt them to provide careless or less thoughtful responses [30, 72].

This limitation has led researchers to turn to alternative sources of data that allow patients to express their perspectives in a more open and unrestricted manner. CRC patient web forums have emerged as a valuable platform for individuals to spontaneously share their experiences, thoughts, and concerns related to CRC treatments [7]. These online communities provide a unique opportunity for patients to engage in discussions about their CRC journey, offering insights into many of their health aspects, including treatment experiences, physical symptoms, emotional well-being, lifestyle adjustments, and social interactions.

As patient web forums continue to provide a wealth of meaningful patient experience information, effectively harnessing and extracting insights from the vast amount of unstructured data within these platforms presents a significant challenge. The sheer volume and diversity of discussions make it impractical for researchers and medical professionals to manually review and analyze each post. This is where the application of natural language processing (NLP) techniques becomes essential.

In recent years, there has been a growing interest in leveraging artificial intelligence (AI) chatbots, such as GPT-3.5 and GPT-4 [49], to assist in the analysis and interpretation of patient forum data. These NLP models have demonstrated remarkable capabilities in language understanding and generation. Together with topic modeling algorithms [40], they can process and analyze large amounts of unstructured text data, extract meaningful topics and generate coherent interpretations based on the given input. However, relying solely on machine-driven interpretations brings certain limitations in colorectal cancer (CRC) research. First of all, topic modeling and generative AI may lack the contextual understanding and domain expertise necessary for accurately interpreting different CRC patient experiences. Furthermore, the reliance on AI-generated interpretations alone can pose a risk of disseminating misleading information, especially when it comes to discovering knowledge on sensitive data, such as publicly available patient experiences, in online CRC patient forums. These limitations underscore the importance of integrating human expertise into the interpretation process of machine intelligence in the CRC research domain.

While AI models offer significant potential in interpreting CRC patient experiences, human experts bring their own unique strengths, such as their professional experience and CRC domain knowledge, to the table. However, human experts also have limitations in their interpretation of patient experiences. Firstly, they have limited time and resources to manually process and interpret large volumes of patient forum data efficiently, especially when it comes to complex tasks like analyzing CRC patient experiences. Moreover, human experts are subject to cognitive limitations, such as information overload and fatigue, which can affect their ability to process and interpret a large amount of online patient discussions accurately. These limitations indicate that human experts clearly require the assistance of

efficient and effective methods to analyze and interpret patient experiences from the large amounts of digital healthcare information available in the CRC research domain.

1.1 Research Questions

In straightforward terms, Dr. Stuart G. Walesh highlights the contrast between humans and machines with the following statement: “The computer is incredibly fast, accurate and stupid. Man is unbelievably slow, inaccurate and brilliant. The marriage of the two is a challenge and opportunity beyond imagination.” [29] In the context of the healthcare domain, it is crucial to close the gap between human and machine interpretation of online colorectal cancer patient forums. Therefore, the following research question is formulated and will be addressed throughout the paper:

RQ. How can topic modeling, GPT-3.5 language generation and human expertise be combined to explore the interpretation of patient web forums in colorectal cancer (CRC) research?

For answering the research question, this thesis aims to compare, combine, and evaluate interpretations generated by both human experts and GPT-3.5 in the context of CRC patient experience topics extracted from patient web forums. To comprehensively explore GPT-3.5 language generation and human expertise, the main research question can be further divided into several in-depth questions:

SQ1. Compare: What are the differences between human and GPT-3.5 in their topic interpretations?

SQ2. Combine: How do human experts reach interpretive agreements with GPT-3.5?

SQ3. Evaluate: What is the added value of combining human and AI topic interpretations in CRC research?

1.2 Thesis Objectives

Throughout this research, three significant objectives will work together to provide answers for the above-mentioned sub-questions:

1. **Comparative Analysis of Human and AI topic interpretations:** The first objective of this research is to conduct a comparative analysis between the topic interpretations generated by human experts and GPT-3.5. This analysis aims to understand how human experts and AI independently interpret the topics related to CRC patient experiences. This serves as a foundational step for further exploration and refinement of the hybrid intelligence approach in the subsequent objectives.
2. **Developing Interpretive Agreements:** The second research objective is to facilitate the development of interpretive agreements by combining the knowledge from human experts and GPT-3.5. This can be achieved through the process of interactive

machine learning [4], where human experts and GPT-3.5 can collaboratively learn from their topic interpretations. By leveraging this interactive process, human experts can continuously refine the AI-generated topic interpretations as well as their own ones based on their personal preferences. Consequently, human experts can decide for themselves when they reach an interpretive agreement: an agreement with GPT-3.5 on the refined human and AI topic interpretations.

3. **Expert Evaluation of Hybrid and Individual Interpretations:** The third and final research objective is to conduct expert evaluations on the individual and hybrid interpretations generated by human experts and GPT-3.5. In this objective, human evaluators, who are the higher experts in the field, assess the relevance of the topic interpretations without prior knowledge of whether they were generated by humans or the AI model. This evaluation process allows for a qualitative assessment of the individual and hybrid interpretations, providing insights into the effectiveness of combining topic interpretations compared to relying solely on individual human and AI interpretations.

1.3 Thesis Outline

This thesis follows the following structure. Firstly, chapter 2 provides a comprehensive review of the relevant literature on the main research components, including topic modeling, GPT-3.5 language generation, and hybrid intelligence. Secondly, chapter 3 presents the methodologies employed throughout the research, covering aspects such as data collection, data analysis, topic interpretation with GPT-3.5 and design-based topic interpretation. Afterwards, chapter 4 addresses *SQ1* by examining the observed differences between individual human and AI topic interpretations. Moreover, chapter 5 focuses on *SQ2*, investigating the methods used for reaching interpretive agreements. Next, chapter 6 answers *SQ3* by analyzing the evaluation results of human experts on hybrid and individual topic interpretations. Finally, chapter 7 concludes the thesis with a summary of the main contributions and outlines future directions for further exploration in this field of research.

Chapter 2

Related Work

The related work chapter explores various aspects related to the study components. It begins by discussing the background of topic modeling, emphasizing its importance and applications within the healthcare domain, while also exploring the assessment methods for the generated topics. Additionally, this chapter covers pre-trained language models, specifically GPT-3.5, and their role in interpreting topic modeling outcomes. Moreover, it mentions background on hybrid intelligence, its diverse applications, and how it can facilitate human involvement in the collaborative interpretation of topic modeling results alongside GPT-3.5. Overall, this chapter plays a crucial role for understanding the methodologies employed in this research.

2.1 Topic Modeling

Topic modeling is a text mining technique that falls under unsupervised machine learning. Its primary purpose is to identify and extract the main themes present within a collection of documents. By analyzing the content of the documents, topic modeling aims to automatically organize and categorize the collection based on the discovered themes [8].

As the healthcare sector continues to generate massive amounts of data, topic modeling has become a valuable approach to extract meaningful information from health and medical corpora [34]. The internet serves as one of the main resources for distributing healthcare data. Online public platforms allow patients and their relatives to learn more about other patients' experiences, including their information needs, communication, and usage of social media for health purposes [12, 21]. Topic modeling can be a valuable method for finding the key health-related topics within uncategorized patient stories on social media websites. Besides that, this method eliminates the need to manually read and analyze extensive volumes of online healthcare information.

2.1.1 Types of Topic Models

There already exist several topic modeling algorithms where each one is developed for its own use case [71]. One of the most popular and widely used topic models is Latent Dirichlet Allocation, which assumes that each document is a mixture of several topics and that each

word in the document is associated with one of the topics [10]. Another topic model is Non-Negative Matrix Factorization, which assumes that the original document-word matrix can be factorized into two non-negative matrices representing document-topic and topic-word relationships [15]. Other algorithms include Hierarchical Dirichlet Process [69], which can automatically learn the number of topics in the data, Latent Semantic Analysis, which uses Singular Value Decomposition to reduce the dimensionality of the term-document matrix [23], and Correlated Topic Model [9], which allows topics to be correlated with each other rather than assumed to be independent.

In recent years, a new approach called BERTopic [33] has gained attention in the field of topic modeling. BERTopic leverages the power of pre-trained language models like BERT (Bidirectional Encoder Representations from Transformers) [26] to generate document embeddings and cluster them into coherent topics. BERTopic offers several advantages, including its ability to capture semantic relationships between words and handle out-of-vocabulary terms effectively using its pre-trained language models.

In light of time constraints and the specific focus of this research, the forthcoming chapters will only explore 3 popular topic models (i.e., Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF) and BERTopic) to discover valuable topics in colorectal cancer patient forums.

2.1.2 Evaluation of Topic Models

Evaluating the goodness of topic modeling outcomes is essential to show that meaningful themes are extracted from the collection of text documents. It is important to note that the study's considered topic models are limited to performing unsupervised tasks, meaning there is no ground truth or predetermined correct topics to compare against. Nevertheless, it is still crucial to validate whether the generated topics are easily understandable and highly interpretable for human reviewers. Three main topic validation methods are particularly relevant to this research:

1. **Topic Coherence:** Since topics may not always be well interpretable, it is important to distinguish good topics from bad ones using coherence measures [64]. This metric quantifies to what extent the highest ranked words within a topic are related to one another. Higher coherence values indicate more meaningful and interpretable topics. It is also widely known that coherence scores can be measured for comparative analysis between different topic modeling methods, such as LDA, NMF and BERTopic, across different numbers of topics [1, 77, 53].
2. **Topic Model Stability:** This metric quantifies the degree of overlap between the topics generated by the model. One possibility is to use the top keywords of each topic pair as input for measuring the Jaccard Similarity Coefficient [56]. The average topic overlap across all pairwise combinations indicates the extent of stability and consistency in the model output. A lower average topic overlap indicates a higher diversity among the topics, meaning that the results are more stable. It is worth to note that standard LDA suffers from very large topic instability due to its non-deterministic behaviour [2].

3. **Human Judgement:** In addition to the automated metrics mentioned earlier, involving human reviewers is essential to ensure that the topic modeling outcome aligns with the intended objectives and requirements of the analysis. Therefore, topic model evaluation should be considered with both automated metrics and human judgement tasks. Existing human metrics in topic modeling often involve direct ratings, as well as word and topic intrusion [35, 20]. These tasks aim to assess the coherence and relevance of the generated topics from a human perspective.

Prior work on human-in-the-loop topic modeling looked at how non-expert end users refined topic models to make them more aligned with user-specific requirements [67]. A good human judgment strategy would be to gain user insights from the perspective of healthcare designers in the context of colorectal cancer. Design literature has shown how data-driven techniques like topic modeling can help with the construction of Patient Community Journey Maps [39, Chapter 3]. It also mentions how designers review the topic information which is provided to them. By involving designers in this research, they can help with identifying and removing poor quality topics generated by the topic models. Therefore, a design-based human judgement approach will be employed for this project, in which a human designer examines the most relevant terms and documents related to each topic as well as the corresponding AI-generated topic interpretation covered in the next section.

2.2 Interpretation with Pre-trained Language Models

Once the topic terms have been extracted for each topic, they can be interpreted by both human experts and machine knowledge. Machine interpretation is facilitated by leveraging pre-trained language models for transforming the topic terms into more interpretable formats. Updating the topic representations can make the topics more readable and understandable for human analysis.

2.2.1 Generative Pre-trained Transformer (GPT)

In the field of natural language processing (NLP), pre-trained language models (PLMs) have become powerful in generating natural language text. With the emergent abilities of large language models (LLMs), it becomes possible to generate coherent and contextually relevant responses to a wide range of prompts to which these models have not been explicitly trained on [73]. One notable LLM is the Generative Pre-trained Transformer (GPT), a deep learning-based model developed by OpenAI¹. GPT is trained on a large corpus of text data from the Internet, including books, articles and web pages [62]. The GPT model releases are listed on the OpenAI platform². The latest model release is GPT-4, which has shown state-of-the-art performance in understanding and generating natural language text or code [57].

¹<https://openai.com/>

²<https://platform.openai.com/docs/models/overview>

2. RELATED WORK

In a previous iteration, OpenAI has developed and released another model named GPT-3.5, also referred to as the brain behind ChatGPT, which still remains to be one of the largest and most complex language models to date. GPT-3.5 is a transformer-based model which has 175 billion parameters, making it much larger than GPT-2, an earlier release with 1.5 billion parameters [61]. Additionally, GPT-3.5 has demonstrated impressive performance on a range of natural language processing tasks, such as language translation, text completion, question-answering, and sentiment analysis [14]. The GPT-3.5 model has the opportunity to grow in several domains, such as scientific research, healthcare and human-machine interaction [36], from which some of them are further discussed in this section.

GPT for Human-Computer Interaction (HCI)

ChatGPT serves as a viable tool for text data analysis in the context of HCI research [68]. Regarding a case with human involvement, OpenAI researchers developed an insightful model called InstructGPT, which is further trained to follow instructions based on human feedback [58]. Consequently, this shows that large language models such as GPT-3.5 possess the capability to enable chatbots in generating more human-like responses for HCI applications [43].

GPT for Healthcare

NLP applications, such as chatbots, can provide technological assistance in analyzing patient data by utilizing Natural Language Understanding (NLU) and Natural Language Generation (NLG) [50]. This additional support equips medical workers with the tools they need to improve patient outcomes in a clinical environment. In the medical domain, GPT-3.5 is useful for many different healthcare services, such as medical decision making, clinical assistance and virtual patient-centered support [38]. Even with the many different potentials of using Artificial Intelligence in healthcare applications, it is unrealistic for GPT-3.5 to replace healthcare providers as they have the necessary human brain power and expertise for making medical decisions [41]. Additionally, the use of GPT-3.5 can lead to pitfalls in healthcare delivery. Some possible misuses of GPT-3.5 in healthcare include the generation of misleading information or the provision of treatment recommendations without proper human assessment [17]. Therefore, integrating GPT-3.5 in medical research and healthcare practice should always happen in the presence of human experts who possess the medical domain knowledge for providing patient care.

2.2.2 Interpreting Topic Models with GPT

In order to make human experts better understand natural language processing outcomes, it is crucial to present the results in a manner that is easily interpretable. For instance, the topics generated by a topic model can be converted to a more representative natural language output using pre-trained transformers. Models like GPT-3.5 have the ability to transform the representation of topic terms by employing zero-shot text classification or text generation tasks specifically designed for topic understanding [31, 47]. Figure 2.1 shows the

work of an existing study which focused on topical language generation, a task that combines topic information from topic modeling with the language generation capabilities of transformer models [76]. To receive representative topic interpretations from GPT-3.5, the prompt pattern should be clearly defined in the input such that the model response is customized according to the user's requirements [74]. This involves prompt-engineering the key topic information with contextualized statements such that GPT-3.5 can clearly describe the topic. Once the topics are interpreted by GPT-3.5, it becomes feasible to compare the results with human interpretations, enabling a comprehensive evaluation from a human standpoint.

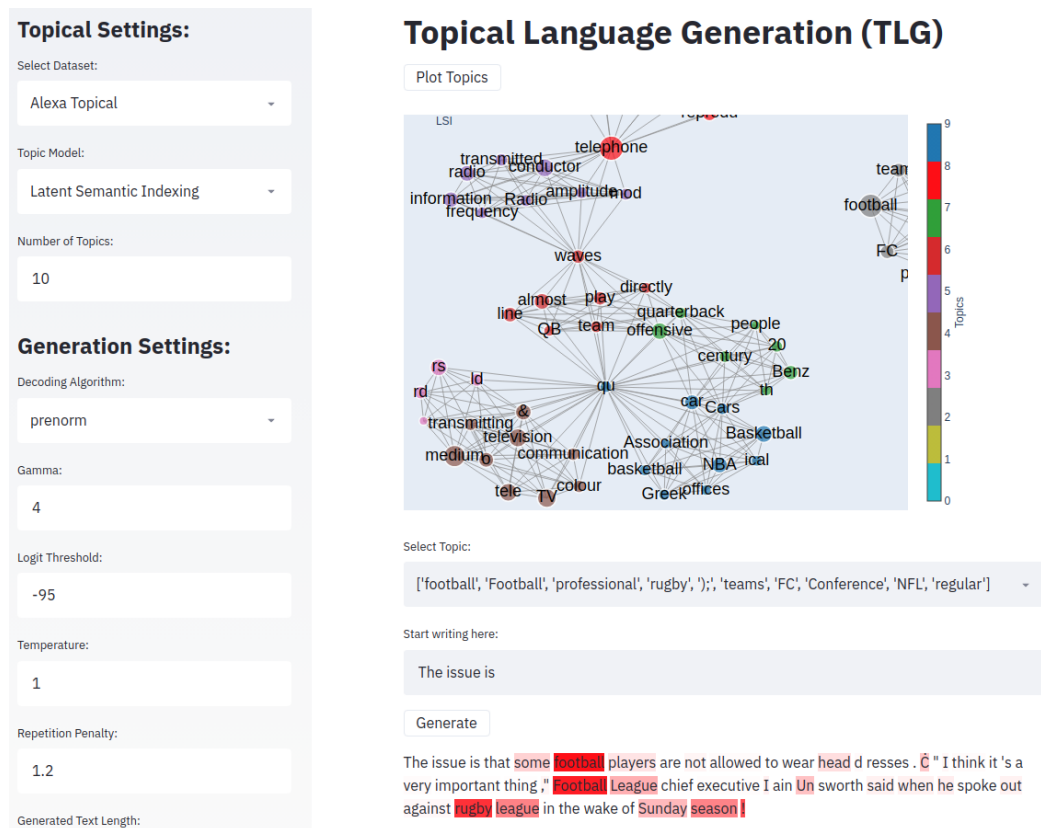


Figure 2.1: Example user interface demonstrating how transformers can generate texts from topic modeling information³.

2.2.3 Evaluating Topic Interpretations

The evaluation of topic interpretations generated by GPT-3.5 makes it necessary to involve healthcare researchers and practitioners in the field of colorectal cancer. By recruiting CRC domain experts to interpret topics and assess the goodness of GPT-3.5 topic interpretation

³https://github.com/roholazandie/topical_language_generation

responses, insights can be gained through comparisons between human-quality text and the output of natural language generation (NLG). There are significant justifications for relying on human judgment instead of automated metrics, such as BLEU [59] and ROUGE [48], to validate the NLG system in the study [65]. First of all, the evaluation is context-dependent, meaning that it is specifically based on colorectal cancer patient experience topics. Second of all, human judgment incorporates domain-specific knowledge by allowing human experts with medical work experience to assess the appropriateness of the GPT-3.5 topic interpretations in the context of colorectal cancer. The best practices for conducting human evaluation of NLG systems typically require good planning, execution and release of the study [70]. A good plan would be to make colorectal cancer researchers analyze and interpret topics by themselves. Thereafter, they are able to provide feedback on the GPT-3.5 generated text by comparing it with their own human-written topic interpretation. In this way, the feedback can be processed by GPT-3.5 to further improve its natural language understanding (NLU) about the topics.

To make things even more interesting and complicated, it turns out that the human evaluation of generated text can also again be evaluated [22]. The study goes beyond simply evaluating GPT-3.5 topic interpretations. According to a recent study, evaluators gave higher ratings to the answers provided by chatbots in response to patient questions compared to those given by physicians [5]. By recruiting evaluators, they can assess the topic relevance of the human-written and GPT-3.5 generated texts. Furthermore, existing work focused on testing whether human evaluators were able to distinguish who created the texts: human or AI [54]. This may also be a good explorative direction for finding aspects that distinguish human-written and AI-generated texts from one another.

2.3 Towards Hybrid Intelligence

Hybrid Intelligence (HI) is all about combining the power of human minds and machine intelligence, to collectively achieve superior results [25]. While the idea of HI sounds promising, research also suggests that AI can have a negative future impact on human society: it can keep growing until it outperforms humans in all areas [60]. This implies that both human and machine have the potential to augment each other's knowledge, depending on the role of the teacher, as illustrated in Figure 2.3 [24, Figure 2]. When it comes to colorectal cancer (CRC) research, the knowledge of generative transformers should be mainly enhanced by human experts, such as researchers and healthcare providers, since they possess a greater depth of CRC domain knowledge compared to AI systems. Therefore, in this study, human expertise plays a crucial role in determining the relevance of patient experiences in topic interpretations. Meanwhile, machine intelligence components, including topic modeling and GPT-3.5, serve as supporting tools for human experts in the process of interpreting patient experience topics related to colorectal cancer.

2.3.1 Hybrid Intelligence Applications

While Hybrid Intelligence can be applied in several domains, it is often implemented in three scenarios: education, healthcare and science [3]. Regarding HI applications in the

healthcare domain, it is important to carefully consider the individual pros and cons that humans and AI bring to the table, as illustrated in Figure 2.2 [19, Figure 2.1]. Healthcare workers and AI must balance each other’s weaknesses by utilizing each other’s strengths. For instance, GPT-3.5 lacks human cognition and CRC domain context, which can be provided by medical experts. On the other hand, the velocity of topic models and GPT-3.5 can reduce the amount of time human experts have to spend in analyzing and interpreting CRC patient information. Thus, it is highly important to combine the strengths of human experts with the ones of GPT-3.5 and topic modeling in CRC research.

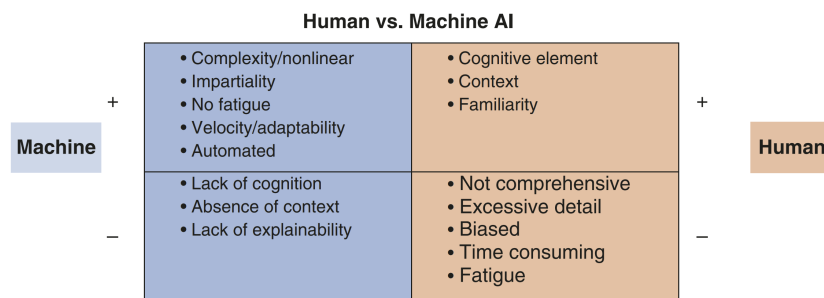


Figure 2.2: Demonstration of the positive and negative aspects of human cognition and artificial intelligence [19, Figure 2.1].

2.3.2 Hybrid Development of Topic Interpretations

There are several justifications for combining human expertise with the pre-trained knowledge of GPT-3.5 to develop hybrid interpretations. Firstly, it facilitates knowledge management [37] between doctors and the AI model, allowing for a mutual exchange and integration of their respective understandings of patient experience topics. This collaborative intelligence increases the overall understanding of the topic at hand. Secondly, this hybrid approach enables human-in-the-loop machine learning [55], since there is an interactive learning process between doctors and GPT-3.5. Human experts can iteratively refine AI-generated interpretations and enhance their own interpretations using insights gained from the AI model. Finally, given that human experts possess greater expertise in healthcare compared to GPT-3.5, it is crucial that the human experts establish trust in the human-AI collaboration process [6] by allowing them to indicate their satisfaction with the refined patient experience topic interpretations.

Similar to performing unsupervised machine learning using topic models, a related study delved into applying hybrid intelligence to generate and evaluate user-specific clusters from financial data [66]. Involving human computation requires setting up a framework where domain experts can interact with the machine to achieve a certain goal. Regarding the interpretation of topics using GPT-3.5 and medical workers, it becomes necessary to facilitate a web application where knowledge sharing can simultaneously take place from both perspectives. With CRC domain experts as the users, they may have the opportunity to collectively develop and refine topic interpretations with the assistance of GPT-3.5. This will lead to topic interpretations which are augmented by both human and AI, following the

2. RELATED WORK

distribution of roles in HI: Augmented Human Intelligence and Augmented Machine Intelligence, as shown in Figure 2.3 [24, Figure 2]. Once the augmented topic interpretations are obtained, these versions and the individual ones can be compared altogether using human evaluation as described in subsection 2.2.3.

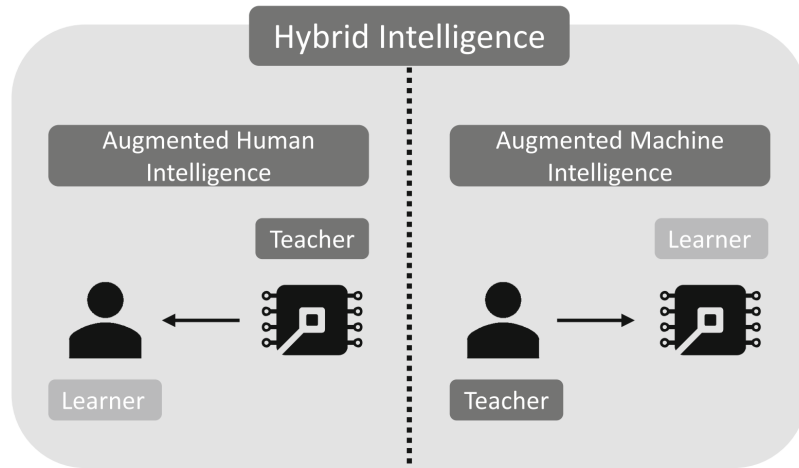


Figure 2.3: Distribution of roles in hybrid intelligence [24, Figure 2].

Chapter 3

Methodologies

Expanding on the previous chapter, the methodologies chapter provides a detailed explanation of how the research components are integrated to form the hybrid intelligence approach in this study. The approach combines topic modeling with GPT-3.5 natural language generation (NLG) and incorporates design-based human judgment in interpreting the discovered topic information.

To begin, the methodology outlines the process of extracting online patient forums and performing topic modeling to discover meaningful CRC patient experience topics. Following the topic modeling process, the methodology discusses how interpretable the identified topics are by evaluating the topic models with automated metrics. Moving forward, the methodology explains how GPT-3.5 is applied to generate natural language text from the extracted topics. Additionally, the methodology highlights topic validation with human judgment by allowing a designer to read and interpret the discovered topic information. Overall, this chapter provides a comprehensive overview of how topic modeling results can be interpreted by GPT-3.5. The understanding of the methodologies is essential for the later chapters, which also involves human experts in the interpretation process of CRC patient experience topics.

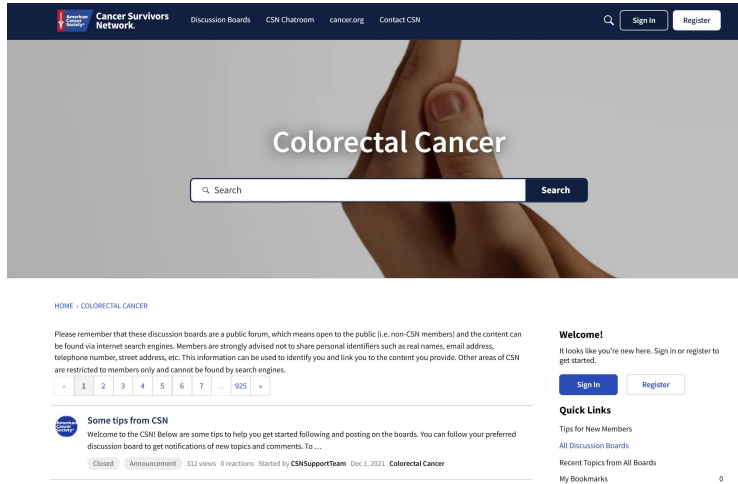
3.1 Data Collection

This study focuses on analyzing colorectal cancer patient forums scraped from the Cancer Survivors Network USA, an open-source patient platform¹. The online platform provides support, education and advocacy for those affected by colorectal cancer, including survivors, caregivers and loved ones. The initial CRC discussion thread has remained active on the public platform since the year 2000. In Figure 3.1a, an overview is provided on the front page of the Cancer Survivors Network USA platform, specifically under the category of colorectal cancer. After navigating to the page of a discussion thread, there are three types of posts which can be encountered. The basic structure of the discussion thread is illustrated in Figure 3.1b. The main post is located at the top, while its corresponding comments are positioned below it. In addition, reply posts may also appear, which contain blockquotes

¹<https://csn.cancer.org/categories/colorectal>

3. METHODOLOGIES

mentioning the post to which they are replying. To minimize duplicate information, the blockquotes are not scraped from the reply posts as their content has already been addressed by earlier posts under the same discussion thread.



(a) CRC frontpage on Cancer Survivors Network USA.



(b) Types of discussion posts under a CRC discussion thread.

Figure 3.1: Brief overview of CRC discussions on Cancer Survivors Network USA.

The CRC posts are extracted from the platform and saved into a local dataset that contains the attributes listed in Table 3.1. The dataset contains more than 290 thousand rows, representing the patient stories that were scraped from the platform at a specific point in time. Therefore, any posts written after the scraping event are not included in the dataset.

url	username	userposts	time	title	post
-----	----------	-----------	------	-------	------

Table 3.1: Cancer Survivors Network USA CRC dataset attributes

The scraping process involves two essential steps: gathering the URLs of discussion pages with the browser automation capabilities of Selenium WebDriver² and parsing the dataset attributes from HTML elements with BeautifulSoup³ while ensuring the security and confidentiality of sensitive information. Personal identifiable information (PII) is removed to protect the privacy of individuals and comply with legal regulations. Medical data, such as colorectal cancer patient information, is highly sensitive, and the leakage of PII can pose significant risks. Therefore, it is crucial to remove attributes like full names, emails, phone numbers, addresses, and URLs from the text to maintain data security.

To achieve this, a combination of techniques is employed. Natural language processing techniques, specifically named entity recognition (NER) [46], are used to identify and classify entities like organizations, locations, names, and events within the text. In this case, the NER functionality from the spaCy library⁴ is utilized. NER helps identify potential PII that needs to be removed. Regular expressions (Regex) [28] are then used to detect and eliminate specific patterns such as international phone numbers, email addresses, postal codes, and URLs.

3.2 Data Analysis

After obtaining the dataset, the next step involves identifying the relevant topics discussed in the colorectal cancer (CRC) patient posts using unsupervised machine learning methods. This process consists of three main stages.

Firstly, the data is pre-processed to enhance its representativeness for subsequent text mining. This pre-processing step helps to transform the raw text into a more suitable format for topic modeling. Secondly, the pre-processed data is explored using different topic modeling techniques, such as Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), and BERTopic. Each technique applies a different set of rules and algorithms to extract topics from the data, as was explained in subsection 2.1.1. Thirdly, for each identified topic, the most relevant topic terms are passed to the generative model GPT-3.5. This model interprets and generates coherent text based on the provided terms, helping to uncover the main themes and content associated with each topic.

3.2.1 Text Cleaning

In the pre-processing phase, several steps are performed to prepare the data for analysis. This subsection outlines the specific techniques applied to clean the text before further analysis. It is important to note that these cleaning steps are only specific to the analysis of LDA

²<https://www.selenium.dev/documentation/webdriver/>

³<https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

⁴<https://spacy.io/>

and NMF. The cleaning steps of BERTopic are performed through its sub-models, which will be explained in section 3.2.2.

- 1. Removal of Punctuations, Stopwords, Usernames and Irrelevant CRC Terms:** Punctuation marks, such as commas, periods, and quotation marks, are removed from the text using regular expressions (Regex). Stopwords are common words that occur frequently in the language but often carry little semantic meaning. Using the English stopwords provided by the Natural Language Toolkit (NLTK), these words, such as ‘the’, ‘and’, ‘is’, are removed from the text. Additionally, usernames from Table 3.1 within the text are also eliminated. Furthermore, in the context of colorectal cancer (CRC) data, specific terms, such as ‘cancer’, ‘doctor’, ‘people’, are too abstract and not so informative to the analysis. Therefore, these irrelevant CRC terms can be identified and removed. Altogether, these removal steps help to eliminate unnecessary noise and special characters that do not contribute significantly to the content analysis.
- 2. Lemmatization of Tokens:** Tokens, or individual words, are lemmatized using the NLTK library. Lemmatization reduces words to their base or dictionary form, such as converting ‘running’ to ‘run’ or ‘better’ to ‘good’. This normalization step helps to increase the frequency of similar words and reduces the overall vocabulary size. To ensure the removal of newly formed stopwords, the lemmatization step is performed before the above-mentioned removal step in item 1.
- 3. Removal of Duplicate Texts:** Duplicate texts within the dataset are identified and removed using the Pandas library’s drop duplicates function⁵. This step ensures that each text instance is unique and prevents redundant information in the analysis.
- 4. Removal of Non-English Texts:** Language detection using the langdetect library⁶ is employed to identify and remove non-English texts from the dataset. This step helps ensure that the analysis focuses solely on English language texts, maintaining consistency and preventing multilingual topics from being formed.

3.2.2 Topic Modeling

This subsection delves into the techniques employed to extract relevant topics from the cleaned patient forum texts using different topic modeling techniques. The three considered techniques are Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), and BERTopic.

Latent Dirichlet Allocation (LDA)

Initially, LDA [10] was chosen as the topic modeling technique due to its speed, scalability, flexibility, and simplicity. However, LDA presented some drawbacks during its implementation. One notable limitation was the significant topic overlap observed in the results. As

⁵https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.drop_duplicates.html

⁶<https://pypi.org/project/langdetect/>

LDA is based on probabilistic mixture rules, the lack of consistent determinism made it challenging to differentiate and interpret distinct topics effectively. LDA treated documents as mixtures of topics, and topics as mixtures of tokens, contributing to the observed overlap.

Determining the ideal number of topics in topic modeling is a critical task. In order to make an informed decision, Figure 3.2 provides a visual representation of the metric lines for various numbers of LDA topics. The configuration setting of the LDA model⁷ included the number of topics, corpus, and id2word dictionary as inputs, while leaving the default model parameters unchanged. Two key automated metrics were used to evaluate the topic models: the average topic overlap measured by Jaccard Similarity [56] and the C_v topic coherence score [64]. The goal was to identify a number of topics that achieved high topic coherence while minimizing topic overlap. The plot demonstrates the relationship between the number of topics and the corresponding values of topic overlap and coherence. Ideally, the chosen number of topics would be where the gap between coherence and topic overlap is maximized. This signifies a configuration that provides highly coherent topics with minimal overlap, ensuring the meaningfulness and distinctiveness of each topic.

Upon closer examination of the plot, it becomes evident that LDA exhibits consistently high levels of topic overlap across the different numbers of topics considered. This observation implies that relying solely on automated metrics may not be sufficient to identify the ideal number of topics. To address this limitation and gain a more comprehensive understanding, a human analysis of the topic terms and their related documents becomes necessary. To overcome these challenges, the focus shifted to alternative techniques, such as Non-Negative Matrix Factorization (NMF) and BERTopic.

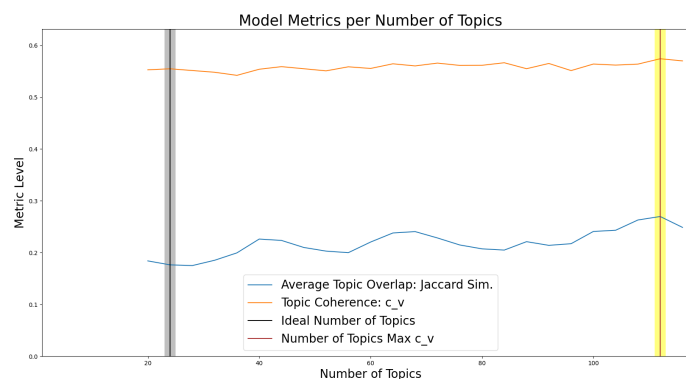


Figure 3.2: Ideal number of LDA topics based on Maximum Difference between Topic Coherence and Topic Overlap (Mean Jaccard Similarity).

Non-Negative Matrix Factorization (NMF)

NMF [15] provided a more deterministic approach to topic modeling where topics represent the linear combinations of words in the corpus. Multiple runs of NMF consistently produced

⁷<https://radimrehurek.com/gensim/models/ldamodel.html>

3. METHODOLOGIES

similar results, allowing for traceability and analysis. Moreover, NMF generated highly interpretable topics with smaller topic overlap compared to LDA based on automated metrics and design-based human judgment, which is described in section 3.4. After a designer performed her developed interpretation strategy to extensively review the most relevant terms and the top 50 highest contributing posts for each topic, it became evident that NMF outperformed LDA in producing clear and understandable topics. Additionally, NMF produced sparser solutions, representing topics with a smaller set of influential words, thus improving conciseness and informativeness. NMF also demonstrated superior efficiency compared to LDA, resulting in faster computations and improved scalability.

In order to demonstrate that NMF exhibits lower topic overlap and higher coherence scores compared to LDA, we present the comparison results of the automated metrics in Table 3.2. The comparison is conducted for different numbers of topics, and the evaluation metrics used include mean Jaccard similarity [56] and Normalized Pointwise Mutual Information (NPMI) coherence score [11]. Both LDA and NMF are run using the same pre-processing steps from subsection 3.2.1, and the top 20 topic terms are considered for computing the metrics. The results clearly indicate that NMF topics exhibit lower overlap, as reflected by the lower mean Jaccard similarity scores. Additionally, for five out of six numbers of topics, NMF topics demonstrate higher coherence scores (C_{NPMI}) compared to LDA. Based on these promising findings, NMF is selected as the preferred choice for the first study, while LDA is primarily utilized for data exploration.

Number of Topics	LDA Coherence	NMF Coherence	LDA Jaccard	NMF Jaccard
20	0.018139	0.0549709	0.1166137	0.0051478
30	0.0201963	0.0404954	0.1119408	0.0024881
40	0.0189193	0.0272445	0.119781	0.0008893
50	0.0133304	0.0138895	0.1579971	0.0005861
60	0.0110088	0.003399	0.1470975	0.0005078

Table 3.2: Comparison between NMF and LDA based on C_{NPMI} coherence score and Average Jaccard Similarity of the top 20 topic terms for different numbers of topics.

NMF baseline. For the first study, a well-suited baseline model was constructed by iteratively experimenting with various automated settings and incorporating a design-based human review process for the NMF topic model. The chosen configuration of this NMF baseline model involved extracting 50 topics from the CRC patient stories. To ensure reproducibility, a random state of 1 was set. The model utilized the Kullback-Leibler [75] beta loss as the objective function and employed the multiplicative update solver [44, Figure 2] for finding an approximate factorization. These specific parameter configurations⁸ were carefully selected to optimize the performance and relevance of the NMF topic model for the study.

⁸<https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html>

BERTopic

In addition to NMF, BERTopic [33] was employed as another topic modeling technique. BERTopic utilizes transformers and c-TF-IDF (class-based TF-IDF) and leverages the power of the BERT model [26] for feature extraction. BERTopic introduces a higher complexity due to its transformer-based architecture. Transformer-based models, such as BERT, have achieved state-of-the-art performance in various natural language processing tasks. BERTopic takes advantage of the advanced capabilities of transformers to uncover complex topics that may be challenging to identify using traditional techniques like LDA or NMF. By leveraging the strengths of transformers and c-TF-IDF, BERTopic is able to extract topics that offer deeper insights into the patient forum texts.

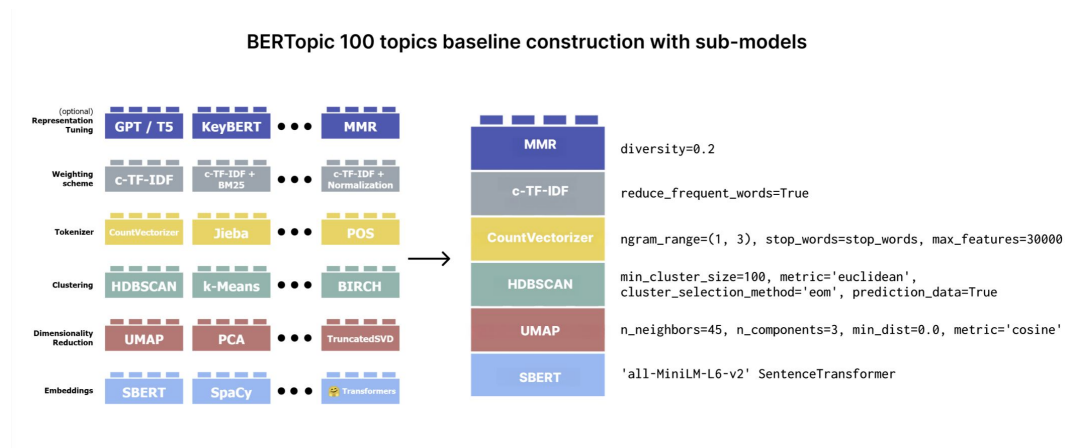


Figure 3.3: BERTopic baseline constructed with 100 topics and the sub-models mentioned on the right side of the arrow.

n_topics	c_v
40	0.38567665307514304
60	0.4286607410347635
80	0.41206038796313643
100	0.4290909570738151
120	0.4300978632827674
140	0.4366577745413183
160	0.42494599326705806
180	0.427550171298985
200	0.4233638702806625
220	0.43177154926104544
240	0.4266092578954601

Table 3.3: BERTopic c_v coherence scores for different numbers of topics using the same model configurations of Figure 3.3.

3. METHODOLOGIES



Figure 3.4: BERTopic visualizations⁹ of the baseline model from Figure 3.3.

By default, BERTopic employs its own approach to extract a suitable number of topics from the provided text data, distinguishing itself from traditional algorithms like LDA and NMF. However, due to the large corpus size, the number of topics that need to be extracted can exceed a thousand. To manage this, a baseline BERTopic model is constructed with a focus on practicality, using 100 topics along with specific submodels and their respective configurations. These details are illustrated in the baseline BERTopic component diagram in Figure 3.3. Topic coherence is not a concern for the BERTopic baseline, since the model consistently generates coherent topics for different runs, as can be observed in Table 3.3. Moreover, the topic visualizations in Figure 3.4 show that the BERTopic baseline makes

⁹https://maartengr.github.io/BERTopic/getting_started/visualization/visualization.html#visualize-topic-similarity

clear distinguishments between the majority of topics, indicating a good stability in the model output.

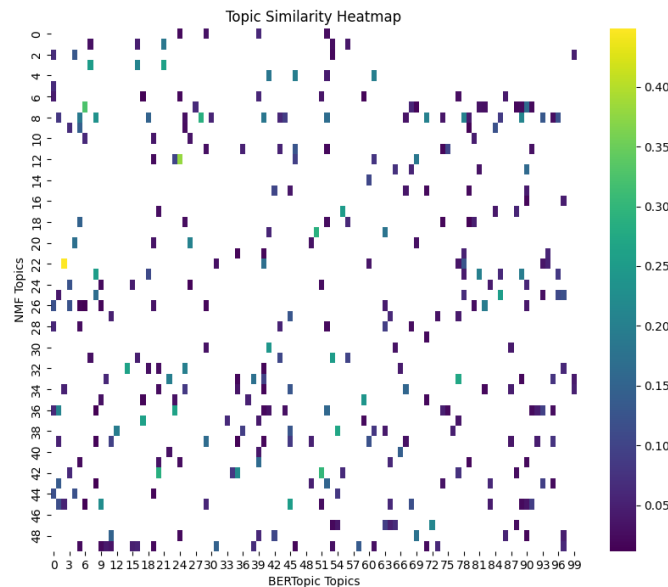


Figure 3.5: Similarity Heatmap between the 50 NMF topics used in the first study and the 100 BERTopic topics employed in the second and third study.

Coming back to the baseline architecture in Figure 3.3, it is worth noting that BERTopic does not require the exact same pre-processing steps mentioned earlier in subsection 3.2.1, which are specific to LDA and NMF. Stopwords and irrelevant CRC terms are excluded by being placed in the ‘stop_words’ parameter of the CountVectorizer model¹⁰. Additionally, the process involves forming bi- and trigrams, as well as reducing the feature space by excluding less frequent terms. HDBSCAN [51] is used to ensure that a minimum amount of patient forum documents are clustered in each topic. Fine-tuning of topic representation is performed using MMR [16], which allows for diversifying the highest ranked keywords and keyphrases. The class-based TF-IDF procedure [33] is applied to reduce the presence of frequently occurring words. For dimensionality reduction, UMAP [52] is used, while Hugging Face’s Mini LM L6 v2 Sentence Transformer¹¹ is responsible for generating embeddings.

While NMF results are utilized for the first study, the inclusion of BERTopic in the second and third studies makes it important to compare the topics generated by the two methods. A comparison between the highest ranking keywords and keyphrases generated by BERTopic and NMF is illustrated in Figure 3.5. Topic vectors are created using sklearn’s TF-IDF vectorizer¹² and the topic overlap between the two methods is measured using

¹⁰https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

¹¹<https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

¹²https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

sklearn’s cosine similarity score¹³. It is observed that the majority of topics from the two baseline models have little to no overlap, as indicated by the purple color in the figure, which corresponds to a similarity score below 0.1. This finding suggests that BERTopic discovers topics that are distinct from those modelled by NMF, covering different aspects of the patient forum data.

3.3 Topic Interpretation with GPT-3.5

The previous section has shown how topic modeling techniques such as LDA, NMF, and BERTopic can be used to extract topics from patient stories. However, the topics are hard to be processed by humans alone due to the lack of readability in topic representations. To tackle this, natural language generation (NLG) tasks can help to transform the topics into more human-readable formats. GPT-3.5, a powerful language model, can achieve this by generating topic interpretations that describe the most relevant terms for each topic, along with the semantic relationships between these terms. Leveraging its language generation capabilities, GPT-3.5 can make it easier for humans to interpret topic modeling outcomes by providing them with interpretations that capture the model’s understanding of the topics. The topic models NMF and BERTopic are relevant to the hybrid intelligence studies of this research, and therefore, it is necessary to interpret the topic terms of these models with pre-trained language models like GPT-3.5.

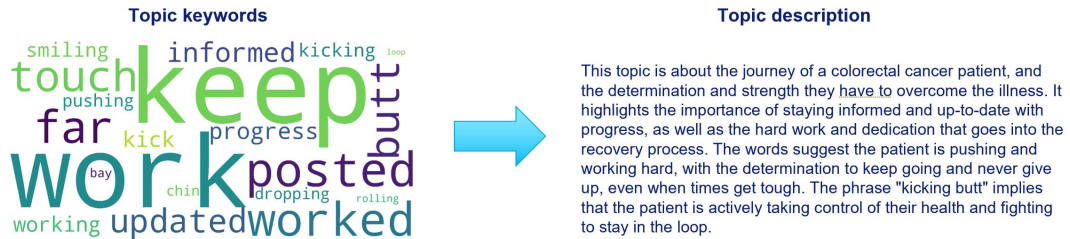
GPT-3.5 can generate meaningful topic interpretations by leveraging a well-defined prompt pattern [74]. For effectively prompt-engineering topic interpretations using GPT-3.5, it is highly important to provide clear objectives and context. The prompt should include background information about the data and the specific topic modeling algorithm used. Additionally, the prompt should explicitly state the task at hand, such as generating a topic interpretation based on the provided topic terms. By incorporating these elements into the prompt, GPT-3.5 can better understand the task and generate more meaningful topic interpretations.

¹³https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html

3.3. Topic Interpretation with GPT-3.5

Engine: 'text-davinci-003', temperature=0.3, top_p=1.0, frequency_penalty=0.0, presence_penalty=0.0, max_tokens=1024

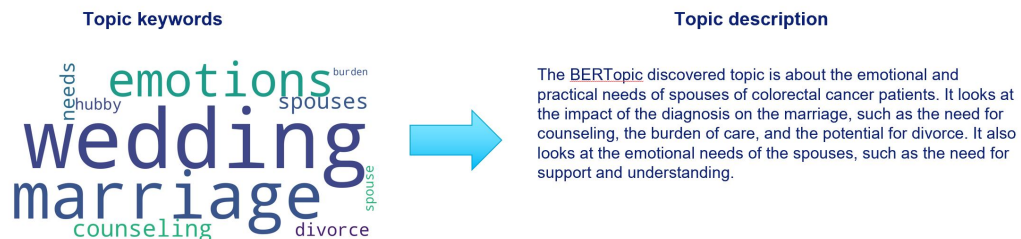
Prompt: I gather online Colorectal Cancer patient stories from Cancer Survivor Network USA and perform NMF topic modelling on the dataset. Explain what the NMF discovered topic is about using its keywords and their corresponding weights: {topic_keywords}. Do not explicitly quote the keywords in the topic description (e.g. "keyword" or "\"keyword\"")!



(a) Example GPT-3.5 Interpretation of one NMF topic.

Engine: 'text-davinci-003', temperature=0.3, top_p=1.0, frequency_penalty=0.0, presence_penalty=0.0, max_tokens=1024

Prompt: I gather online Colorectal Cancer patient stories from Cancer Survivor Network USA and perform BERTopic topic modelling on the dataset. Explain what the BERTopic discovered topic is about using its keywords and their corresponding weights: {keyword_weights}. Do not explicitly quote the keywords in the topic description (e.g. "keyword" or "\"keyword\"")!



(b) Example GPT-3.5 Interpretation of one BERTopic topic.

Figure 3.6: Examples of Topic Interpretation Generation by the 'text-davinci-003' GPT-3.5 Model Release. The significance of topic terms in the word cloud is reflected by their respective sizes, with larger terms indicating higher relevance to the topic.

In addition to improving the prompt text, customizing the model response also relies on fine-tuning the model parameters. Controlling the length of the generated text is necessary for avoiding excessively long model responses while still enabling the model to provide sufficient information about the topic. Furthermore, the randomness of the generated text can be easily controlled using the model temperature. Keeping the temperature at a low value like 0.3 is crucial for making the model response more deterministic and focused on accurate topic interpretations. Figure 3.6 presents examples on how OpenAI's Completions API¹⁴ is employed to generate topic interpretations for NMF and BERTopic results.

¹⁴<https://platform.openai.com/docs/guides/gpt/completions-api>

3.4 Design-based Topic Interpretation

Aside from topical language generation with GPT-3.5, the topic modeling results can also be interpreted and understood using human knowledge. To facilitate this process, the highest ranked terms and the 50 most relevant documents associated with each topic are provided to an Industrial Design student from TU Delft. The student thoroughly reviews the topic information and provides human judgment for each topic from a designer's perspective.

Regarding the designer's in-depth interpretations on the NMF baseline, which was mentioned in section 3.2.2, it was discovered that among the 50 NMF topics, 12 primarily revolved around online activities within the Cancer Survivors Network USA forum. However, these forum-related topics did not contribute any significant value specifically to colorectal cancer (CRC) patient stories. The forum-related topics primarily consisted of basic information, such as sending prayers and condolences, as well as celebrating positive news. Therefore, these topics were excluded from the NMF results for the purpose of the first study.

To further analyze the remaining 38 NMF topics, the designer examined the topic descriptions generated by GPT-3.5. Next to the most relevant terms and posts for each topic, these Natural Language Generation (NLG) responses served as an additional source of topic information before she provided her own interpretation for each topic. During this process, the designer distinguished between topics focused on hospital patient experiences, indicated by the color blue, and those centered around home patient experiences, indicated by the color orange. In addition, Figure 3.7 shows the upper clusters she created to represent the topics that belong to the same patient experience category. For the first study, the clusters and their corresponding topics were presented in the form of a Patient Journey Map, as detailed in Appendix A. This map presented a user-friendly design visualization, which enables the study participants to easily identify the specific stage of the cancer journey to which the topics and clusters were applicable.

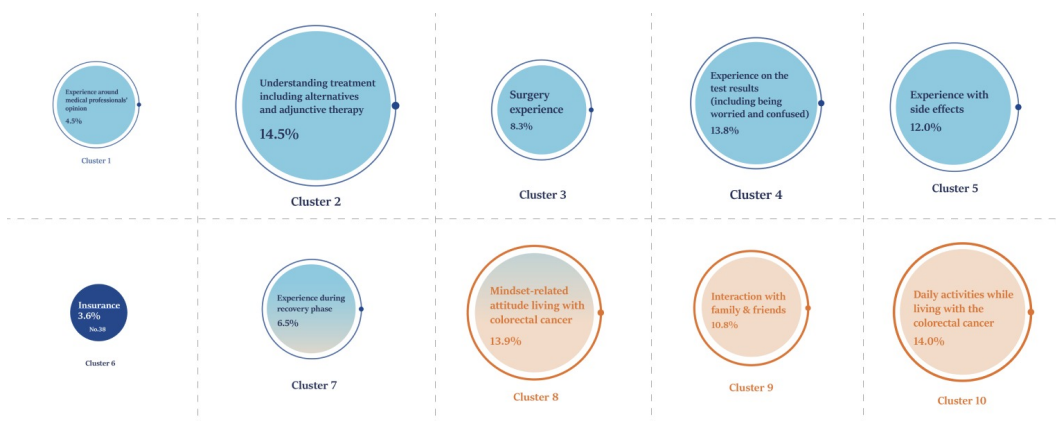


Figure 3.7: The NMF generated upper clusters along with their respective sizes indicating how many documents are contained within each cluster.

For the purpose of the second and third studies, the designer also reviewed the BERTopic

baseline, which was shown in Figure 3.3. In analyzing these results, the designer identified that 21 out of the 100 topics were either related to the forum or lacked insights regarding CRC patient experiences. To identify these irrelevant topics, the designer reviewed the top 10 keyphrases, the top 50 most relevant posts, and the GPT-3.5 NLG response for each topic.

After the first study, the focus had been placed on mainly interpreting the BERTopic results with medical expertise and GPT-3.5 language generation. Therefore, the designer did not extensively analyze the topics generated by BERTopic unlike the comprehensive interpretation applied to the NMF topics. Table 3.4 offers a summarized overview of the tasks performed by the designer during the analysis of the NMF and BERTopic baselines.

NMF	BERTopic
Reviewed top 20 keywords for each topic	Reviewed top 10 keyphrases for each topic
Reviewed top 50 posts for each topic	Reviewed top 50 posts for each topic
Reviewed GPT-3.5 interpretation for each topic	Reviewed GPT-3.5 interpretation for each topic
Identified 12 out of 50 topics as redundant	Identified 21 out of 100 topics as redundant
Wrote down designer interpretations for the remaining topics	
Generated upper clusters for the remaining topics	
Classified the remaining topics to ‘home’ and ‘hospital’	
Constructed a Patient Journey Map from the remaining topics and the generated clusters	

Table 3.4: Summary of tasks performed by the designer for the NMF and BERTopic baselines.

Chapter 4

Comparative Analysis of Human and AI Interpretations

While the previous chapter explored topic modeling and the topic interpretations generated by GPT-3.5 and a human designer, it lacks the valuable input from medical workers specializing in CRC research. This input is crucial given the focus of this research on analyzing the experiences of CRC patients. By allowing human experts to interpret topics, one can draw comparisons between their interpretations and those of GPT-3.5, resulting in observed differences which clarify the individual approaches in which humans and GPT-3.5 interpret topics. These findings will provide an answer to *SQL*. *What are the differences between Human and GPT-3.5 in their topic interpretations?*

The chapter is structured into two main parts. Firstly, the study procedure is outlined, detailing how medical workers interpret NMF topics and compare their interpretations with those of GPT-3.5. Secondly, the study findings are discussed, highlighting the differences observed between human and GPT-3.5 interpretations through a comparative analysis of their derivation methods.

4.1 Study Procedure

In the first study, a comparative analysis was carried out to differentiate the topic interpretations generated by the AI model GPT-3.5 and the ones provided by medical workers. The focus of the analysis was on the NMF baseline results from section 3.2.2.

In collaboration with the designer, we recruited 5 medical workers specializing in colorectal cancer healthcare. Together, we organized a co-creation session during which the medical workers were given the chance to interpret NMF topics and provide feedback on the AI-generated topic interpretations.

4.1.1 Participant Work Experience

At the beginning of the co-creation session with the medical workers, we administered an ‘Expert Experience Survey’ to validate their level of expertise in the CRC domain, ensuring their ability to understand and interpret patient experiences accurately.

4. COMPARATIVE ANALYSIS OF HUMAN AND AI INTERPRETATIONS

Figure 4.1 provides an overview of the participants' experience and their reliability to participate in the study. The figure illustrates that the five participants consist of two surgeons, two Ph.D. candidates, and one epidemiologist. Notably, three out of the five participants possess more than ten years of work experience, indicating a high level of expertise in the healthcare domain.

The survey responses indicate that the participants have limited familiarity with machine learning and are not acquainted with ChatGPT. Consequently, this study, which involves NMF topic modeling and AI topic interpretations, presents a novel experience for them.

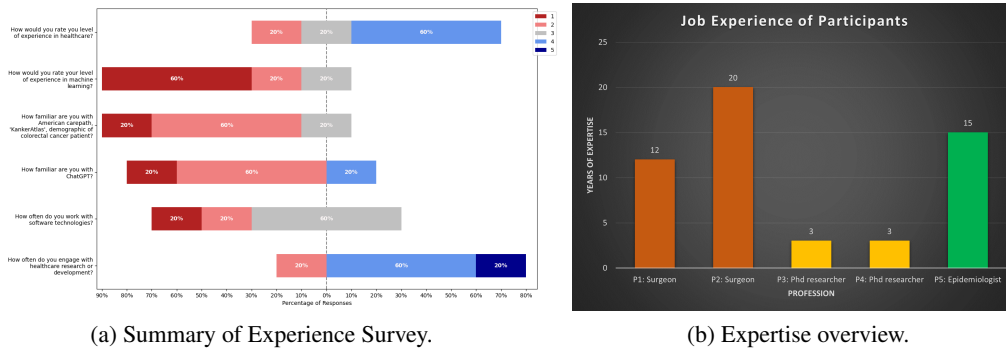


Figure 4.1: Summary of Participant Work Experience.

4.1.2 Participant Task

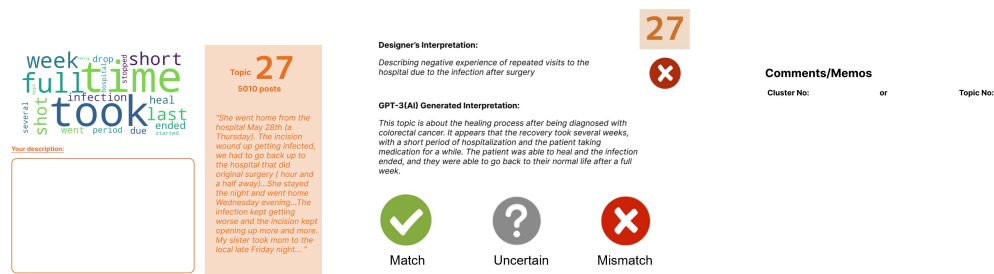
During the co-creation session, we engaged the 5 medical workers in an activity where they were provided with 10 NMF cluster cards displaying the different topic clusters from Figure 3.7, to which the 38 NMF baseline topic cards were classified. The participants were asked to select at most two cluster cards that they found most significant to their field of expertise. Once the cluster cards were chosen, the participants were tasked with interpreting and validating the topic cards within those selected clusters.

To begin the interpretation process, the participants were given the opportunity to familiarize themselves with the most relevant NMF topic terms, which were ordered based on their weights of importance in a topic word cloud. Additionally, they were provided with a quote extracted from one of the top 50 relevant topic posts that had been reviewed by the designer in relation to the NMF results. Subsequently, the participants were instructed to write down their topic interpretation on the front side of the topic cards. An example of a topic card's front side is showcased in Figure 4.2a.

Afterwards, the participants engaged in comparing their topic interpretations with those generated by GPT-3.5. To ensure a fair comparison, the AI-generated interpretations were kept hidden on the back side of each topic card until this point in the activity. Notably, the back side of the card also included the designer's interpretation, serving as a reference to the participants during this comparison process. An example of a topic card's back side is showcased in Figure 4.2b. During this phase, the participants had the opportunity to verify

whether their interpretation matched, did not match, or if they were uncertain about the interpretation provided by GPT-3.5.

The co-creation session reached its conclusion by providing the participants with comment cards which made them share their feedback on the topic interpretations generated by GPT-3.5. The template of these comment cards can be seen in Figure 4.2c. The purpose of this feedback was to provide participants with an opportunity to express their agreement or disagreement with the AI-generated topic interpretations.



(a) Front-side of a NMF Topic Card. The participant can read the topic terms and the topic quote and write a topic description on the card.

(b) Back-side of a NMF Topic Card. The participants can read the topic interpretations of the designer and GPT-3.5 and compare whether these ones match with theirs.

(c) Comment Card where the participants can write down their comments on GPT-3.5 topic interpretations.

Figure 4.2: Example of a topic card and a comment card used for human expert interpretations on NMF topics during the co-creation session of study 1.

4.2 Study Reflection

At the end of the co-creation session, we collected the topic and comment cards from the five participants. Topics from different cluster cards were evaluated by participants with diverse areas of expertise. To account for this variation, the evaluated topics were summarized for each participant. Subsequently, we delve into the main observations that differentiate the approaches between human and GPT-3.5 in interpreting the NMF topic modeling results.

4.2.1 Participant Evaluation Summary

To begin the reflection, we initially concentrated on understanding the primary interests of each participant by examining the cluster cards they chose. Additionally, we investigated the topics they reviewed to assess the performance of each participant in interpreting the NMF topic modelling results.

The cluster cards selected by the five participants are presented in Table 4.1. It is evident that the participants demonstrated interest in only 5 out of the 10 available cluster cards from Figure 3.7. Analyzing the table reveals that three out of the five participants were particularly focused on understanding the experiences of patients regarding test results. Additionally, it is worth noting that clusters related to the daily activities and mindset-related

4. COMPARATIVE ANALYSIS OF HUMAN AND AI INTERPRETATIONS

attitudes of CRC patients were selected twice, indicating a higher level of interest in these topics. On the other hand, clusters associated with the side-effects and recovery experiences of CRC patients were chosen only once, suggesting comparatively less focus on these particular aspects.

Participant	Selected Cluster Card
P1: Surgeon	Cluster 4: Experience on the test results (including being worried and confused)
P1: Surgeon	Cluster 8: Mindset-related attitude living with colorectal cancer
P2: Surgeon	Cluster 7: Experience during recovery phase
P2: Surgeon	Cluster 10: Daily activities while living with colorectal cancer
P3: Researcher	Cluster 4: Experience on the test results (including being worried and confused)
P4: Researcher	Cluster 4: Experience on the test results (including being worried and confused)
P4: Researcher	Cluster 10: Daily activities while living with colorectal cancer
P5: Epidemiologist	Cluster 5: Experience with side-effects
P5: Epidemiologist	Cluster 8: Mindset-related attitude living with colorectal cancer

Table 4.1: Selected clusters for each participant.

Moving forward, the participants were able to interpret and comment on topics within each selected cluster. A comprehensive overview of the evaluated topics is provided in Figure 4.3. This overview includes the number of topics interpreted by each participant, the number of comments made by each participant on AI topic interpretations, and the count of topics where participants indicated a mismatch between their own interpretations and those generated by AI. Notably, the first three participants exclusively provided feedback on topics where there was a mismatch between their own interpretations and the AI-generated interpretations. On the other hand, the last two participants commented on all of their interpretations, including both matched and mismatched ones. The full sets of human-written topic cards and comment cards from the 5 study participants can be observed in Appendix B.

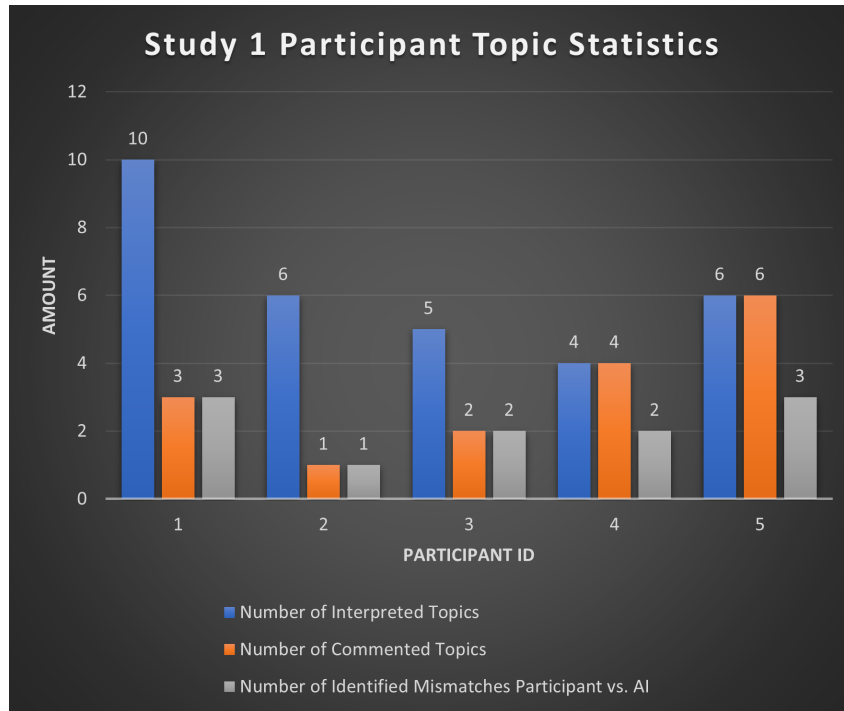


Figure 4.3: Summary of evaluated topics per participant.

4.2.2 Study Outcome

Following the review of participants' expertise information and the comparison of their topic interpretations with those generated by GPT-3.5, several key observations emerge that differentiate the two types of interpretations.

Observed Human Behavior

The interpretation of human experts relies on understanding CRC patient experience use cases within the applicable topic. Unlike simply presenting experts with a list of words for each NMF topic, the inclusion of patient stories allows them to grasp the context surrounding the topic and gain a deeper understanding of patients' perspectives. By providing full sentences from the topic posts, patient scenarios become understandable to the experts, making it easier for them to interpret what patients express in the forum posts. As an example, Participant 3, a researcher, generated Cluster 4's human interpretations that heavily describe the patients' perspectives from the topic quotes. The topic card interpretations of this participant can be observed in Figure B.5.

Another notable observation is that human interpretations are dependent on the professional experience of the human expert involved. Based on their unique experiences and expertise, participants are able to recognize the primary patient needs or concerns from the topic cards. For example, Participant 1, a surgeon specializing in CRC, consistently interprets the topics within Cluster 4 by emphasizing the themes of anxiety, fear, and un-

4. COMPARATIVE ANALYSIS OF HUMAN AND AI INTERPRETATIONS

certainty related to test results. Given surgeons' regular interaction with CRC patients and their expertise in treating the condition, it is understandable that they can easily recognize the emotional experiences patients undergo when seeking further information about their treatment findings. Figure B.1 presents the topic card interpretations written by Participant 1 for Cluster 4.

Another core observation distinguishing humans from GPT-3.5 in their topic interpretations is subjectivity. Different individuals can describe the same topic in various ways based on their unique perspectives and biases. This observation is possible, because Table 4.1 has demonstrated that different participants selected the same clusters. As evident from Figure B.4 and Figure B.7, Participant 2 (i.e., a surgeon) and Participant 4 (i.e., a researcher) offer different interpretations for Topic Card 31 from Cluster 10. The surgeon interprets the topic as *picking up "normal life"*, whereas the researcher perceives it as *celebration of life for cancer survivors*. This example highlights that human understanding of a topic can vary based on individual subjective opinions and areas of expertise. Therefore, individual perspective biases are apparent in human interpretations, while GPT-3.5 generates its interpretations using its own pre-trained knowledge, without the influence of personal preferences.

Observed AI Behavior

Regarding the behavior of GPT-3.5 topic interpretations, it can be observed that the AI model interprets topics by employing its natural language processing capabilities to recognize the main patterns and associations among topic terms. This allows GPT-3.5 to describe the connections between specific topic terms more easily. As a pre-trained language model, GPT-3.5 possesses the advantage of efficiently processing large amounts of information, enabling it to identify associations between topic terms more efficiently than manual human analysis. For example, GPT-3.5 demonstrates the ability to recognize that the topic terms within Cluster 5 primarily indicate the emotional and physical side-effects associated with CRC. This understanding is reflected in the interpretations generated by the AI model, as presented in Figure B.8.

Furthermore, inconsistent patient stories make it more challenging for GPT-3.5 to capture the main topic theme. In such cases, the AI-generated interpretation tends to align better with human interpretations when it primarily focuses on describing the most relevant topic terms rather than attempting to incorporate the diverse patient use cases. This suggests that GPT-3.5 can be easily influenced by patient stories with different contexts. For example, Figure 4.4 shows that GPT-3.5 covers various aspects of topic 49 while interpreting different variations of topic posts. The topic cards from Participant 1 and 3 can be fully observed in Figure B.1 and Figure B.5. The human interpretations tend to be more focused on the emotional aspects of CRC patient experience, whereas the AI-generated interpretation with five topic posts does not mention any patient concerns. Therefore, by providing more diverse topic information to generate an interpretation, GPT-3.5 tends to deviate from primarily focusing on the topic terms.

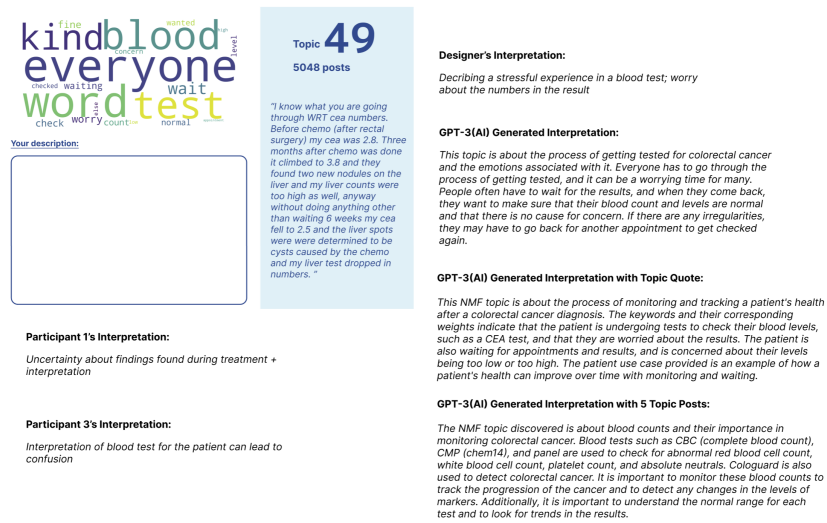


Figure 4.4: Example observation of GPT-3.5's behavioral change for interpreting diverse topic information.

A final remarkable observation distinguishing GPT-3.5 from humans is that the pre-trained language model does not inherently make ethical considerations when interpreting patient stories. Patient stories consist of sensitive information and should be correctly understood by both human and AI. However, it has been observed that GPT-3.5 can sometimes misinterpret patient experiences based on the topic terms alone. For example, in the case of Topic 27 from Cluster 7, Participant 2 (a surgeon) and the designer classified the topic under the treatment phase of the CRC patient journey. On the other hand, GPT-3.5 interprets the topic as a patient experience during the diagnosis phase. The treatment and diagnosis classifications can be observed in the designer's patient journey map in Appendix A, as well as in Participant 2's Topic 27 Comment Card shown in Figure B.3. This observation suggests that GPT-3.5 can provide misleading topic information when analyzing patient data, and therefore, human experts cannot solely depend on AI models for sensitive medical data analysis.

4. COMPARATIVE ANALYSIS OF HUMAN AND AI INTERPRETATIONS

Unique Observation	Observation Description	Example Scenario
<i>A. Human Observations</i>		
Human's Contextual Understanding	Human Intelligence typically requires understanding the context present in patient use cases.	P3 (researcher) consistently describes the patients' perspectives from the topic quotes
Human's Domain Expertise	Human Intelligence is capable to integrate experts' domain knowledge into the topic interpretation.	P1 (surgeon) consistently interprets anxiety, fear and uncertainty based on his professional practice
Human's Subjectivity	Human Intelligence can offer patient experience insights from different human perspectives.	P2 (surgeon) describes topic 31 as <i>picking up "normal life"</i> , whereas P4 (researcher) describes the same topic as <i>celebration of life for cancer survivors</i>
<i>B. AI Observations</i>		
AI's Pattern Recognition	AI uses NLP to recognize the patient experience from the topic terms.	GPT-3.5 recognizes that cluster 5 consists of topics about emotional and physical side-effects of CRC patients
AI's Behavioral Change for Inconsistent Patient Stories	Different patient stories make it challenging for AI to capture the main topic theme.	GPT-3.5 deviates from interpreting the topic terms from topic 49 when the AI model also has to interpret diverse topic posts
AI's Missed Ethical Considerations	Sensitive information, such as CRC treatment and diagnosis, can be misinterpreted by AI on its own.	P2 interprets topic 27 as a patient experience during the treatment phase, whereas GPT-3.5 interprets it as a patient experience during the diagnosis phase

Table 4.2: Scenario-based observations that differentiate how human experts and GPT-3.5 interpret NMF topic modeling results.

Chapter 5

Developing Interpretive Agreements

While the previous chapter focused on comparing the interpretations of humans and AI from their respective standpoints, this chapter takes a step further by exploring the potential of combining the knowledge of CRC human experts and GPT-3.5 to develop hybrid topic interpretations. Through knowledge sharing on CRC patient experience topics, human experts can refine the individual human and AI topic interpretations. This collaborative process aims to increase the understanding of CRC patient experiences and aims to develop agreements on the interpretations created by human and AI. By investigating how human experts form hybrid topic interpretations with the assistance of GPT-3.5, the following research question will be answered: *SQ2. How do Human Experts reach interpretive agreements with GPT-3.5?*

This chapter is structured as follows. Firstly, it will discuss the web application method which allows human experts to engage with GPT-3.5 on CRC patient experiences, facilitating the development of hybrid topic interpretations. Secondly, the study findings will be presented, highlighting the methods which make human experts establish interpretive agreements with GPT-3.5.

5.1 Study Procedure

For the purpose of this study, a total of 12 medical researchers from Erasmus MC were recruited to participate in a dedicated session, during which they were provided with an opportunity to interpret the BERTopic baseline results from section 3.2.2. Subsequently, they engaged in an interactive process with GPT-3.5 to refine both human and AI topic interpretations. To make their activities possible, a Python Flask application was implemented [32]. This application served as a platform to deliver the essential topic information and interactive instructions, thereby enabling the participants to collaborate effectively and develop interpretive agreements with GPT-3.5.

5. DEVELOPING INTERPRETIVE AGREEMENTS

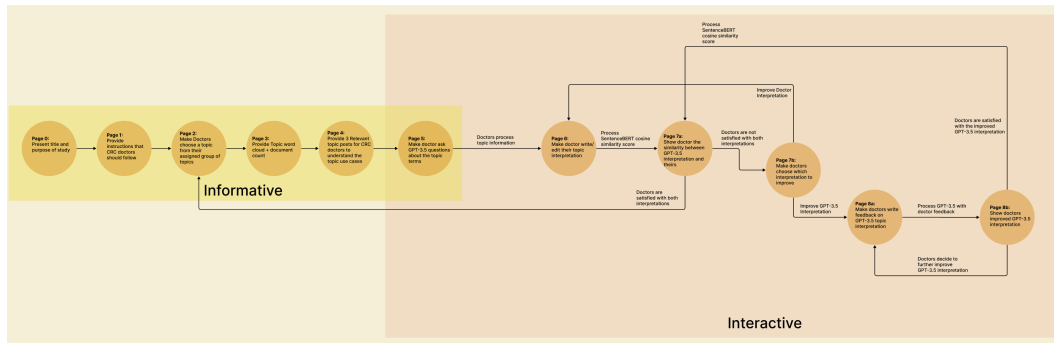


Figure 5.1: Flow diagram for combining GPT-3.5 and doctor interpretations through Flask Web Application.

The entire flow diagram of the Flask application is presented in Figure 5.1, while Appendix C offers a more detailed overview of the specific Flask web pages developed for the study. The web application consists of two main components: an informative part which provides users with the necessary topic information, and an interactive part which engages users to refine both human and GPT-3.5 topic interpretations. The sequential steps outlining the flow of the application are explained below.

- (Page0.) **Title and Study Goal:** The users were presented with the title and purpose of the study as shown in Figure C.1a.
- (Page1.) **User Instructions:** In Figure C.1b, the users were provided with an overview of tasks they had to follow.
- (Page2.) **Selecting a Topic:** To ensure that every topic was interpreted at least once, each participant was assigned a predetermined group of topics derived from the BERTopic baseline. The users had to fill in their Participant ID on the page shown in Figure C.1c. Afterwards, they were able to access the topics they were assigned on the page presented in Figure C.1d. Upon completing all tasks related to a specific topic, participants were redirected back to the page on Figure C.1c for further engagement with subsequent topics.
- (Page3.) **Reading Topic Keyphrases:** Similar to the co-creation session in the first study, the users were presented with a topic word cloud and document count for each topic, as can be seen in Figure C.1e. These visual representations showcased the top 10 keyphrases associated with the topic as well as the number of forum posts belonging to that particular topic.
- (Page4.) **Reading Topic Posts:** To increase the understanding for each topic, the page on Figure C.1f provides the users with three patient stories extracted from the top 50 most relevant posts associated with that specific topic. Compared to the topic quote utilized in the first study, this expanded approach provides human experts with additional patient use cases. This offers them the opportunity to analyze more patient information for which the topic is applicable.

(Page5.) **Asking Topic Questions to GPT-3.5:** As an optional source of topic information, the page on Figure C.1g gave users the opportunity to ask topic-related questions to GPT-3.5. They were allowed to ask general questions or questions specifically related to their selected keyphrases. These keyphrases were then used in the prompt of GPT-3.5, allowing for context-aware responses that align with the user interests. Figure 5.2 presents an example of how this question answering task is prompt engineered.

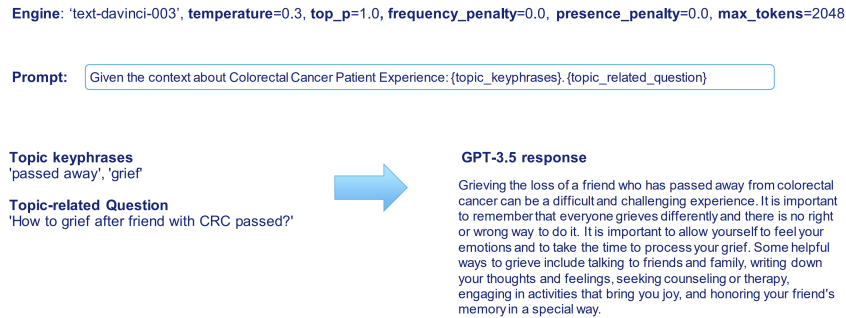


Figure 5.2: Zero-shot Question Answering task to get Topic Information from GPT-3.5.

(Page6.) **Writing a Topic Description:** After reviewing and processing all of the provided topic information, participants were required to share their understanding of the CRC patient experience topic by writing a topic description on the page illustrated in Figure C.1h. This task involved combining their insights and knowledge gained from the topic word cloud, document count, topic posts, and any additional topic information extracted from GPT-3.5.

(Page7.) **Comparing Human and GPT-3.5 Descriptions:** Upon submitting their human description, participants were directed to page 7 shown in Figure C.1i. This page allowed the users to review both Human and AI interpretations. At the top of the page, users could observe the list of topic keyphrases for reference. Moreover, both the human-written and GPT-3.5-generated topic interpretations were displayed side by side, enabling a direct comparison between the two. Additionally, Semantic Textual Similarity (STS) can give users an idea of how well two blocks of texts align in terms of semantic equivalence [18]. Therefore, a percentage value was also provided, indicating the semantic similarity between the two interpretations. This measure was computed by transforming the texts into vector embeddings using the 'paraphrase-MiniLM-L6-v2' model¹ from the SentenceTransformers framework [63], and then calculating the cosine similarity score² between the embeddings, which was converted into a percentage representation.

After reviewing the interpretations, users were presented with three options. They could choose to refine their own description by editing their text on page 6 or refine

¹<https://huggingface.co/sentence-transformers/paraphrase-MiniLM-L6-v2>

²https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html

5. DEVELOPING INTERPRETIVE AGREEMENTS

the GPT-3.5 description by proceeding to page 8. If users agree with both interpretations, they could express their satisfaction by clicking the green button on this page. This action would redirect the users back to page 2 and save the initial human and GPT-3.5 interpretations, along with the final human and GPT-3.5 interpretations, for the corresponding topic. Overall, this page allowed users to actively decide how to refine the topic interpretations based on their assessment and preferences.

(Page 8.) **Writing Feedback on GPT-3.5 Description:** Upon selecting the option to refine the GPT-3.5 description, users were brought to page 8 as shown in Figure C.11. At the top of the page, users could provide their feedback on the initial GPT-3.5 description through a dedicated text box. Directly below, two side-by-side boxes were displayed, with the left box containing the original GPT-3.5 description and the right box showcasing the refined GPT-3.5 description. The incorporation of user feedback into the GPT-3.5 topic interpretation became possible with prompt-engineering, as illustrated in Figure 5.3.

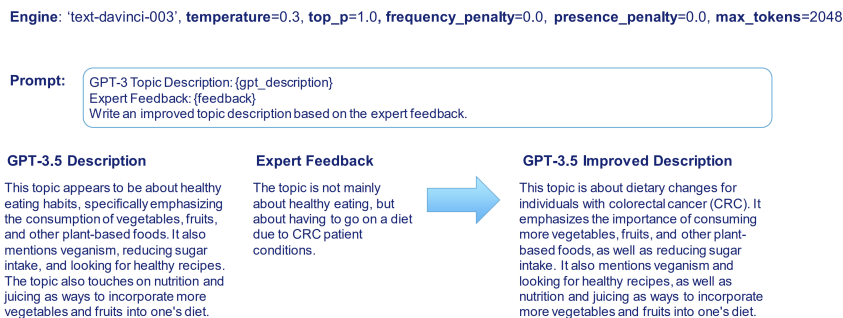


Figure 5.3: Zero-shot text-to-text transformation task to refine GPT-3.5 topic description based on Expert Feedback.

After the users completed the iterative process of refining the initial GPT-3.5 description, they had the option to submit their improvements and proceed back to page 7. It is important to note that users were only able to refine the original description and could not directly modify the newly improved version. If users wished to further improve the updated description, they were required to submit the current version and come back to page 8 for additional refinements.

After the participants' completion of interpreting their pre-selected topics in collaboration with GPT-3.5 on the Flask web application, all of their topic interpretation outputs were locally stored in JSON format. Table 5.1 shows the attributes stored for each topic assessed by the participants. These attributes will collectively provide insights on how human experts assess and refine topic interpretations with the assistance of GPT-3.5.

Attribute	Description
session_id	Flask session ID which uniquely identifies the participant
topic_id	The specific topic evaluated by the participant
human_description	Original human interpretation
gpt_description	Original AI interpretation
improved_human_description	Final improved human interpretation
improved_gpt_description	Final improved AI interpretation
similarity_score	Similarity score between final human and AI interpretations
response_time	Total duration for reviewing and interpreting the topic in collaboration with GPT-3.5

Table 5.1: Saved attributes for each topic assessed by the participants.

5.2 Study Reflection

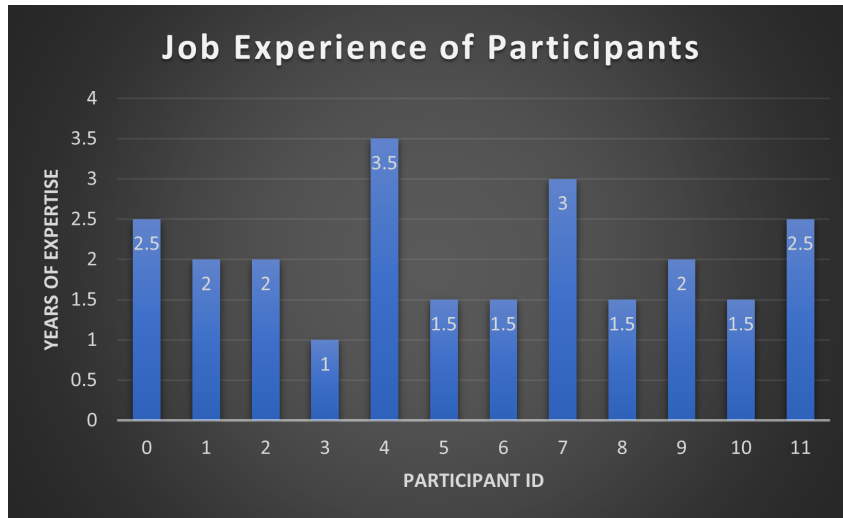
At the end of the human-computer interaction phase with the 12 medical researchers, they were requested to fill out an expert experience survey to gather feedback on their work experience and Flask user experience. Additionally, they were given the opportunity to share additional user experiences through recordings at the end of the study. Firstly, this section will provide an overview of the contributions made by each participant in this study, including their expertise and user experiences on the web application. This will be followed by the study outcomes, which will explain the main methods that led to the participants reaching interpretive agreements based on their collected interpretation outputs.

5.2.1 Participant Evaluation Summary

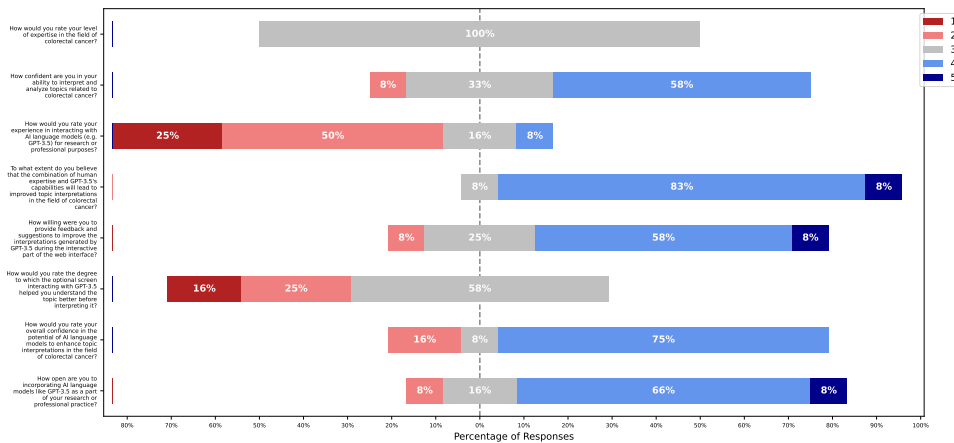
Similar to the first study, we carefully considered the level of experience of the medical participants to ensure that they possessed the necessary domain knowledge to effectively interpret CRC patient experiences. At the time of the study, all 12 participants were highly qualified Ph.D. researchers at Erasmus MC. Each participant possessed at least one year of expertise in the healthcare domain, as presented in Figure 5.4a. In addition, all 12 participants considered themselves to have a sufficient level of expertise in the field of CRC, as shown in Figure 5.4b.

According to the other findings presented in Figure 5.4b, the participants indicated a high level of confidence in interpreting CRC patient experience topics based on their Flask user experience. While most participants had limited prior experience in interacting with GPT-3.5, they still believe that human and GPT-3.5 knowledge can be combined to enhance the understanding of patient experiences in the field of CRC. Furthermore, the majority of participants expressed openness to incorporating AI models into their research or professional practice. Overall, the participants were mostly willing to integrate AI in healthcare and to explore the collaborative potential of hybrid intelligence in CRC research.

5. DEVELOPING INTERPRETIVE AGREEMENTS



(a) Expertise Overview.



(b) Summary of Experience Survey.

Figure 5.4: Summary of Participant Expertise and Flask User Experience.

Moving forward to the remaining user experiences from Figure 5.4b as well as the ones mentioned in the recorded feedback of the participants, they reflected on the following aspects more specifically focused on the Flask web application pages.

- 1. Instruction Page:** The instructions from Figure C.1b were clear and helped the participants in understanding their tasks.
- 2. Topic keyphrases:** The topic word cloud was generally clear to understand, but there were instances of overlap or lack of significance within the keyphrases.

3. **Topic posts:** Three topic posts provided some information, but varied in terms of relevance. This made it challenging to summarize them into a single topic description. It was suggested that more posts might be beneficial for certain topics to understand a broader range of patient use cases.
4. **GPT-3.5 Optional Screen:** Some participants used the optional screen to ask about the meaning of specific topic keyphrases, while others felt that more topic posts would be preferable over this additional source of topic information. In most cases, this optional screen was not used at all, since participants could already understand the topic from the word cloud and topic posts.
5. **Human vs. AI Comparison Screen:** Participants found the listing of topic keyphrases on the comparison screen to be helpful, but the suggestion was made to also make them appear while writing the topic description. Furthermore, the similarity score had an influence on the participants' interpretation behavior. Some participants focused on improving the score until it reached a satisfactory level.
6. **GPT-3.5 Feedback and Improvements:** GPT-3.5 feedback was not always processed as expected. Multiple attempts were often needed to improve interpretations based on the feedback. Some participants had to provide detailed comments for GPT-3.5 to understand the required improvements. Nevertheless, they were still willing to iteratively improve the GPT-3.5 interpretations with their feedback.

Overall, the participants made some usability suggestions (e.g. navigation improvements and clearer topic information presentation), but still had a positive and successful experience while interpreting topics and interacting with GPT-3.5 to form hybrid interpretations. Participants were able to recognize that the topics were based on different types of patient experiences, such as the emotional, social and physical aspects of patients. And most importantly, they expressed confidence in their ability to accurately refine human and AI interpretations according to their personal preferences.

5.2.2 Study Outcome

In order to examine how the participants achieved interpretive agreements with GPT-3.5, an analysis of the collected attributes from Table 5.1 is conducted. This analysis involves comparing the original and improved human and AI interpretations to determine the refinements made by the participants in developing hybrid interpretations. By examining these comparisons, insights can be derived regarding the specific refinements made to achieve agreements between human and AI-generated interpretations.

From the 100 BERTopic baseline results, 112 sets of topic interpretations were produced by the 12 participants, meaning that some topics were interpreted more than once. This amount was reduced to 89 by excluding interpretations of the 21 topics that were previously identified as redundant by the designer in Table 3.4 and deemed unnecessary for the analysis of participants' interpretations.

The distribution of how participants reached interpretive agreements with GPT-3.5 can be categorized into four main methods, as illustrated in Figure 5.5. The figure highlights

5. DEVELOPING INTERPRETIVE AGREEMENTS

that the most frequently occurring cases were those where participants aimed to solely refine the topic interpretations generated by GPT-3.5, followed by cases where participants only focused on refining their own human interpretations. In only 17% of all cases, participants were directly satisfied after their interpretation and immediately reached an interpretive agreement without the need for any refinements. Overall, the 89 sets of topic interpretations are well-distributed over the four categories, covering enough data samples for further analysis within each type of interpretive agreement.

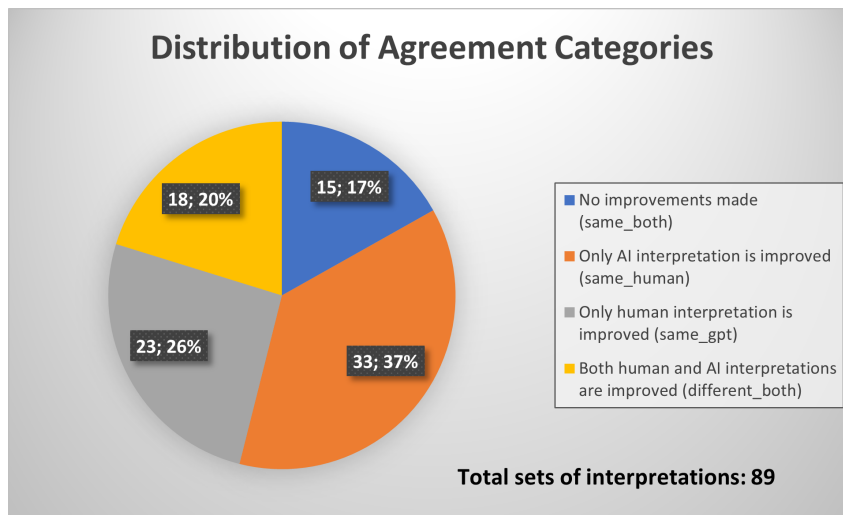
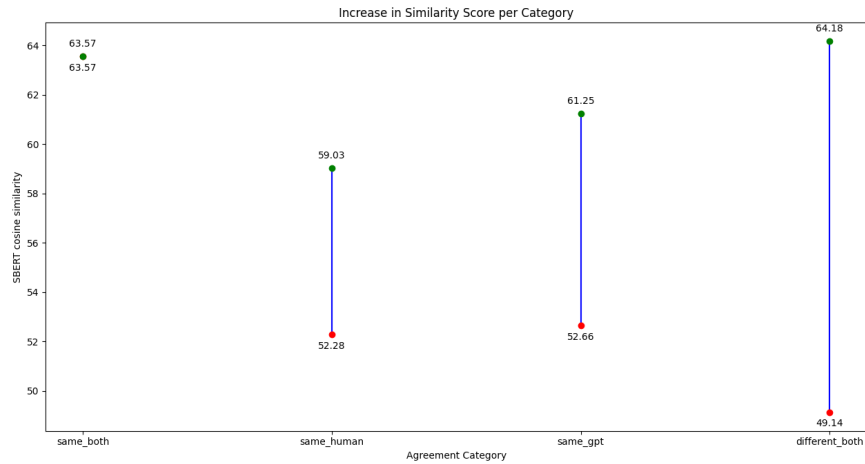


Figure 5.5: Distribution of Methods for Reaching Interpretive Agreements.

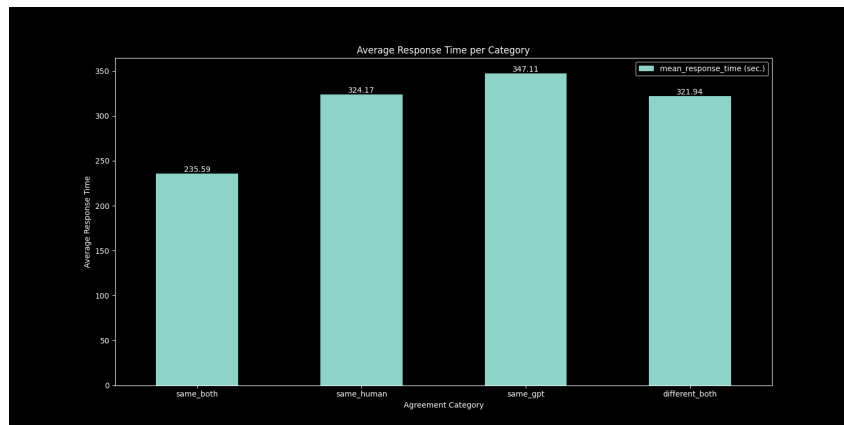
Figure 5.6a presents the average percentage increase in Human-AI similarity for reaching each interpretive agreement category. It showcases the improvement in agreement between human and AI interpretations from the initial topic interpretations to the final hybrid interpretations. As observed in the figure, the 'same_both' category shows no increase as participants did not make any changes to the initial interpretations. The 'same_human' and 'same_gpt' categories show a moderate level of similarity increase, since participants focus on improving either their own initial interpretation or the AI-generated interpretation, but not both simultaneously. On the other hand, the 'different_both' category exhibits the highest level of similarity increase, suggesting that participants achieved the most significant improvements by modifying both human and AI topic interpretations. This finding implies that refining and aligning both sources of interpretation led to a more effective convergence between human and AI perspectives.

Figure 5.6b displays the average total response time for reaching each interpretive agreement category, including the time required to review and interpret the topic information. The 'same_both' category has the shortest response time, suggesting that participants did not make any refinements and were already satisfied with the initial human and AI interpretations. For the other categories, participants also had to spend time adjusting their initial topic interpretation, providing feedback to improve the AI-generated interpretation, and carefully comparing and evaluating the refined interpretations to determine their level

of agreement with GPT-3.5. Therefore, this additional effort and time spent in refining and assessing the interpretations contributed to a longer response time compared to the ‘same_both’ category, where no refinements were made.



(a) Mean Similarity Increase per Agreement Category.



(b) Mean Response Time per Agreement Category.

Figure 5.6: Human-AI Similarity and Response Time Measurements for each type of Interpretive Agreement.

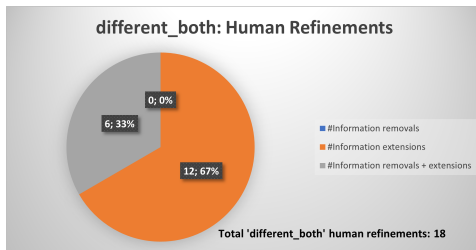
How to reach interpretive agreements with GPT-3.5?	#Cases / #Total (Percentage)	Initial average Human-AI similarity (%)	Final average Human-AI similarity (%)	Similarity increase (%)	Response time (s)
No refinements made (same_both)	15/89 (17%)	63.57	63.57	0	235.59
Only AI interpretation is refined (same_human)	33/89 (37%)	52.28	59.03	6.75	324.17
Only Human interpretation is refined (same_gpt)	23/89 (26%)	52.66	61.25	8.59	347.11
Both Human and AI interpretations are refined (different_both)	18/89 (20%)	49.14	64.18	15.04	321.94

Table 5.2: Summary of Numerical Analysis of Interpretive Agreement Methods.

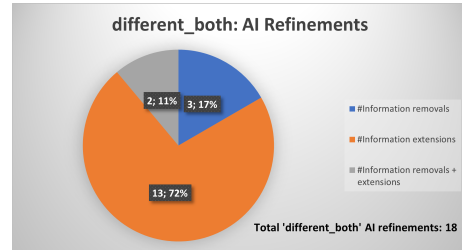
5. DEVELOPING INTERPRETIVE AGREEMENTS

A summarized overview of the numerical analysis within each method of interpretive agreement is presented in Table 5.2. This table reflects on the previous discussions regarding the agreement distribution and the performance measurements, such as human-AI similarity and response time, for each type of interpretive agreement.

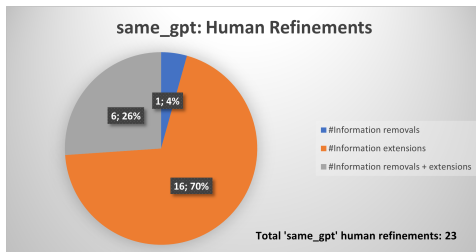
The distributions of how participants approached their refinements in human and AI topic interpretations are presented in Figure 5.7. Analysis of the refinements made to the original interpretations revealed a notable trend: both human and AI consistently extended the information in the majority of cases. This indicates that both human and AI primarily focused on adding new value to the topic interpretations. The objective was to provide a clearer and more comprehensive description of the colorectal cancer (CRC) patient experience topic involved. By primarily expanding topic information with one another, both human and AI versions of the interpretations aimed to increase the understanding and relevance of the topics to the context of CRC patient experiences.



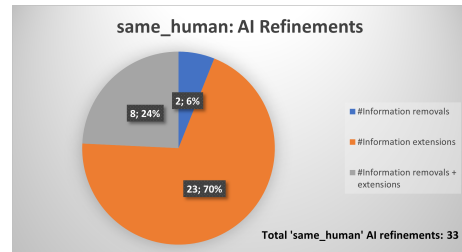
(a) Human refinements made in 'different_both'.



(b) AI refinements made in 'different_both'.



(c) Human refinements made in 'same_gpt'.



(d) AI refinements made in 'same_human'.

Figure 5.7: Distributions of refinement methods by human experts on human and AI interpretations.

Figure 5.8 presents the different patient contexts in which human and AI refinements were made. These patient experiences were recognized by clustering the topics after reviewing the refined human-written and AI-generated interpretations. The full sets of saved interpretations for each refinement category can be observed in Appendix D. By analyzing the patient context for these refined interpretations, it becomes possible to identify the CRC patient experiences for which the participants disagreed with GPT-3.5 and vice versa. Therefore, the exploration of the four possible methods for reaching interpretive agreements are summarized as follows:

5.2. Study Reflection

AI refinements	Tasks (AI, Human)	Subject	Example (incl. task nr.)	Legend:
same_human				
Refinements	21	Physical activities to manage colorectal cancer symptoms	21 Refinements: 'Physical activities to manage colorectal cancer symptoms' and 'Physical activities to manage colorectal cancer symptoms'.	<ul style="list-style-type: none"> Physical activities to manage colorectal cancer symptoms Diagnostic tools for cancer detection Implications for detecting lung nodules Patients' negative emotional experiences Expressing support for colorectal cancer Patients seeking advice about their mental and physical conditions Influence of lifestyle habits on treatment outcomes Motivation and discipline of colorectal cancer patients Relationship between chemotherapy and its side-effects Calibrating Milestones Physical activities to manage colorectal cancer symptoms Physical effects of colorectal cancer
Extensions	26, 30, 35, 36, 45, 46, 48, 49, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100	<ul style="list-style-type: none"> Relationship between chemotherapy and its side-effects Distraction techniques for colorectal cancer patients Physical effects of colorectal cancer Patients' negative emotional experiences Relationship between chemotherapy and its side-effects Calibrating Milestones Physical activities to manage colorectal cancer symptoms Physical effects of colorectal cancer 	<ul style="list-style-type: none"> 20) Details with 'bassins and washing of a new side effect of chemotherapy' 81) Details with 'CRC treatment plan is given at home or outside, allowing for greater flexibility and convenience' 91) Details with 'a new option to help manage the treatment plan' 92) Details with 'importance of critical evaluation on medical treatments and diagnostic tests' 93) Details with 'many results related to diagnosis and treatment of various diseases and conditions' 94) Details with 'types of cancer diagnosis' 95) Details with 'types of cancer diagnosis' 96) Details with 'types of cancer diagnosis' 97) Details with 'types of cancer diagnosis' 98) Details with 'types of cancer diagnosis' 99) Details with 'types of cancer diagnosis' 100) Details with 'types of cancer diagnosis' 	
Refinements + Extensions	22, 37, 40	<ul style="list-style-type: none"> Relationship between chemotherapy and its side-effects Distraction techniques for colorectal cancer patients Physical effects of colorectal cancer Patients' negative emotional experiences Relationship between chemotherapy and its side-effects Calibrating Milestones Physical activities to manage colorectal cancer symptoms Physical effects of colorectal cancer 	<ul style="list-style-type: none"> 22) Refinements: 'Relationship between chemotherapy and its side-effects' and 'Relationship between chemotherapy and its side-effects'. 37) Refinements: 'Distraction techniques for colorectal cancer patients' and 'Distraction techniques for colorectal cancer patients'. 40) Refinements: 'Physical effects of colorectal cancer' and 'Physical effects of colorectal cancer'. 	

(a) Patient Context - AI refinements of 'same_human'.

Human refinements	Tasks (AI, Human)	Subject	Example (incl. task nr.)	Legend:
same_gpt				
Refinements	21	Relationship between chemotherapy and its side-effects	21 Refinements: 'Relationship between chemotherapy and its side-effects' and 'Relationship between chemotherapy and its side-effects'.	<ul style="list-style-type: none"> Treatment options for colorectal cancer (before and after cancer recurrence) Diagnostic tools for cancer detection Patients' negative emotional experiences Methods for monitoring colorectal cancer Patients' financial losses of having colorectal cancer Influence of lifestyle habits on treatment outcomes Motivation and discipline of colorectal cancer patients Relationship between chemotherapy and its side-effects Calibrating Milestones Physical activities to manage colorectal cancer symptoms Physical effects of colorectal cancer Patients' memories
Extensions	2, 4, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100	<ul style="list-style-type: none"> Relationship between chemotherapy and its side-effects Distraction techniques for colorectal cancer patients Physical effects of colorectal cancer Patients' negative emotional experiences Relationship between chemotherapy and its side-effects Calibrating Milestones Physical activities to manage colorectal cancer symptoms Physical effects of colorectal cancer 	<ul style="list-style-type: none"> 80) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 81) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 82) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 83) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 84) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 85) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 86) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 87) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 88) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 89) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 90) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 	
Refinements + Extensions	14, 15	<ul style="list-style-type: none"> Relationship between chemotherapy and its side-effects Distraction techniques for colorectal cancer patients Physical effects of colorectal cancer Patients' negative emotional experiences Relationship between chemotherapy and its side-effects Calibrating Milestones Physical activities to manage colorectal cancer symptoms Physical effects of colorectal cancer 	<ul style="list-style-type: none"> 14) Refinements: 'Relationship between chemotherapy and its side-effects' and 'Relationship between chemotherapy and its side-effects'. 15) Refinements: 'Distraction techniques for colorectal cancer patients' and 'Distraction techniques for colorectal cancer patients'. 	

(b) Patient Context - Human refinements of 'same_gpt'.

Human refinements	Tasks (AI, Human)	Subject	Example (incl. task nr.)	Legend:
different_both				
Refinements	80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100	<ul style="list-style-type: none"> Relationship between chemotherapy and its side-effects Distraction techniques for colorectal cancer patients Physical effects of colorectal cancer Patients' negative emotional experiences Relationship between chemotherapy and its side-effects Calibrating Milestones Physical activities to manage colorectal cancer symptoms Physical effects of colorectal cancer 	<ul style="list-style-type: none"> 80) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 81) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 82) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 83) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 84) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 85) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 86) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 87) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 88) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 89) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 90) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 	
Refinements + Extensions	14, 15	<ul style="list-style-type: none"> Relationship between chemotherapy and its side-effects Distraction techniques for colorectal cancer patients Physical effects of colorectal cancer Patients' negative emotional experiences Relationship between chemotherapy and its side-effects Calibrating Milestones Physical activities to manage colorectal cancer symptoms Physical effects of colorectal cancer 	<ul style="list-style-type: none"> 14) Refinements: 'Relationship between chemotherapy and its side-effects' and 'Relationship between chemotherapy and its side-effects'. 15) Refinements: 'Distraction techniques for colorectal cancer patients' and 'Distraction techniques for colorectal cancer patients'. 	

(c) Patient Context - Human refinements of 'different_both'.

AI refinements	Tasks (AI, Human)	Subject	Example (incl. task nr.)	Legend:
different_both				
Refinements	80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100	<ul style="list-style-type: none"> Relationship between chemotherapy and its side-effects Distraction techniques for colorectal cancer patients Physical effects of colorectal cancer Patients' negative emotional experiences Relationship between chemotherapy and its side-effects Calibrating Milestones Physical activities to manage colorectal cancer symptoms Physical effects of colorectal cancer 	<ul style="list-style-type: none"> 80) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 81) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 82) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 83) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 84) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 85) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 86) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 87) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 88) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 89) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 90) Details with 'having related issues in patients with colorectal cancer who are treated with chemotherapy' 	
Extensions	14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100	<ul style="list-style-type: none"> Relationship between chemotherapy and its side-effects Distraction techniques for colorectal cancer patients Physical effects of colorectal cancer Patients' negative emotional experiences Relationship between chemotherapy and its side-effects Calibrating Milestones Physical activities to manage colorectal cancer symptoms Physical effects of colorectal cancer 	<ul style="list-style-type: none"> 14) Refinements: 'Relationship between chemotherapy and its side-effects' and 'Relationship between chemotherapy and its side-effects'. 15) Refinements: 'Distraction techniques for colorectal cancer patients' and 'Distraction techniques for colorectal cancer patients'. 	
Refinements + Extensions	14, 15	<ul style="list-style-type: none"> Relationship between chemotherapy and its side-effects Distraction techniques for colorectal cancer patients Physical effects of colorectal cancer Patients' negative emotional experiences Relationship between chemotherapy and its side-effects Calibrating Milestones Physical activities to manage colorectal cancer symptoms Physical effects of colorectal cancer 	<ul style="list-style-type: none"> 14) Refinements: 'Relationship between chemotherapy and its side-effects' and 'Relationship between chemotherapy and its side-effects'. 15) Refinements: 'Distraction techniques for colorectal cancer patients' and 'Distraction techniques for colorectal cancer patients'. 	

(d) Patient Context - AI refinements of 'different_both'.

Figure 5.8: Summary of Human and AI refinements for different agreement categories.

- same_both:** Participants being satisfied with the initial human and AI interpretations. In this category, no refinements could be made according to the participants. Figure 5.6a shows that this category has a mean similarity of 63.57%, meaning that the original human and AI interpretations already lied close to each other in terms of semantic equivalence.
- same_human:** Participants choosing to solely refine the AI-generated interpretation. In most of the cases of Figure 5.8a, the participants primarily focused on enhancing the understanding of GPT-3.5 with chemotherapy side-effects and with additional information on CRC treatment options. This implies that GPT-3.5 may not possess the same level of 'treatment' expertise as medical researchers, thereby requiring human

5. DEVELOPING INTERPRETIVE AGREEMENTS

intervention to ensure a comprehensive interpretation of patient experiences during the treatment phase. In general, these participants had confidence in writing their own topic interpretations, as indicated in Figure 5.4b.

3. **same_gpt:** Participants choosing to solely refine their own human interpretation. Figure 5.8b shows that participants consistently refined the same popular patient experiences: relationship between chemotherapy and its side-effects, as well as treatment options for colorectal cancer. This clearly implies that, even in their field of ‘treatment’ expertise, the participants overlooked certain aspects of the CRC treatment and acknowledged the value of AI-generated interpretations as a reference point to enhance their own understanding.
4. **different_both:** Participants deciding to refine both the human and AI interpretations. The process of knowledge integration in this category primarily increased the amount of information in the human and AI topic interpretations, as can be seen in Figure 5.7a and Figure 5.7b. Refining both human and AI interpretations led to the highest average increase in semantic equivalence, as can be seen in Figure 5.6a, indicating that knowledge sharing had the highest impact when both human and AI took part in the interactive engagement. Furthermore, Figure 5.8c and Figure 5.8d show that the interests of the participants mainly lied in refining both human and AI interpretations with CRC treatment options, alternative CRC treatment options and distraction techniques for CRC patients. This suggests that their decision to interchange information on these types of patient experiences with GPT-3.5 explored the full potential of this hybrid intelligence approach by combining the individual strengths of human and AI. They integrated their domain expertise by describing what they understood on CRC treatments and CRC patients’ distraction activities, while keeping their own interpretation up-to-date with the help of GPT-3.5. This two-sided refinement process indicates that there is a high interest for medical researchers to enhance the understanding of CRC patient experiences in the areas of treatment options, alternative treatments, and strategies for emotional well-being.

Chapter 6

Expert Evaluation of Hybrid and Individual Interpretations

After establishing interpretive agreements in the previous chapter, the process of combining human expertise with GPT-3.5 to develop hybrid interpretations on CRC patient experience topics became evident. However, one last essential step is to conduct human evaluations to assess the individual and hybrid topic interpretations and determine their relevance to their respective topics. This evaluation aims to provide an answer to the following research question: *SQ3. What is the added value of combining human and AI topic interpretations in CRC research?*

This chapter is divided into two main sections. The first section focuses on the evaluation method used to assess the relevance of individual and hybrid human and AI topic interpretations. The second section presents the findings of the study, examining whether hybrid interpretations demonstrate superior topic relevance compared to individual interpretations.

6.1 Study Procedure

In the third study, the expertise of 3 full-time medical professionals was harnessed to evaluate the topic relevance of individual and hybrid human and AI interpretations on CRC patient experiences. These medical professionals, who have extensive work experience in the field, were invited to participate in an evaluation process by filling out a dedicated evaluation form.

Due to time constraints, a subset of topic interpretations was selected for evaluation in the third study. This subset consisted of interpretations where the refinements made by human experts and AI had significantly altered the semantic meaning of the original descriptions. To identify these topics, the sets of interpretations from the previous study were sorted based on the difference in similarity increase between human and AI interpretations, as can be seen in the figures of Appendix D. The subset of topics chosen for evaluation were those that had undergone substantial semantic changes in their refined versions. This approach ensured that the evaluated interpretations represented cases where the human and

6. EXPERT EVALUATION OF HYBRID AND INDIVIDUAL INTERPRETATIONS

AI contributions had resulted in noticeable differences in the topic descriptions.

The subset of interpretations selected for evaluation in the third study consisted of 27 out of the possible 89 sets of topic interpretations. These sets of interpretations were collected by drawing 9 sets of topic interpretations from each refinement category: ‘same_human’, ‘same_gpt’ and ‘different_both’. It is worth to note that these selected sets of interpretations were the ones that exhibited significant semantic changes in the second study. In the ‘same_human’ and ‘same_gpt’ categories, where either human or AI interpretations were refined, a total of three topic interpretations were evaluated: two original and one refined in each category. In the ‘different_both’ category, where both human and AI interpretations were refined, a total of four topic interpretations were evaluated: two original and two refined. This ensured a comprehensive evaluation of the different refinement methods and their impact on the interpretations.

For the evaluation process in the third study, an online survey tool called LimeSurvey¹ was used to create an evaluation form. This form enabled human evaluators to review the topic information and provide ratings for the individual and hybrid interpretations generated in the previous study. Each of the 3 human evaluators was assigned with 3 sets of interpretations in round-robin fashion per refinement category, resulting in the total of 9 sets of interpretations being evaluated by each participant. The entire process of constructing the 3 interpretation evaluation surveys can be observed in Figure 6.1

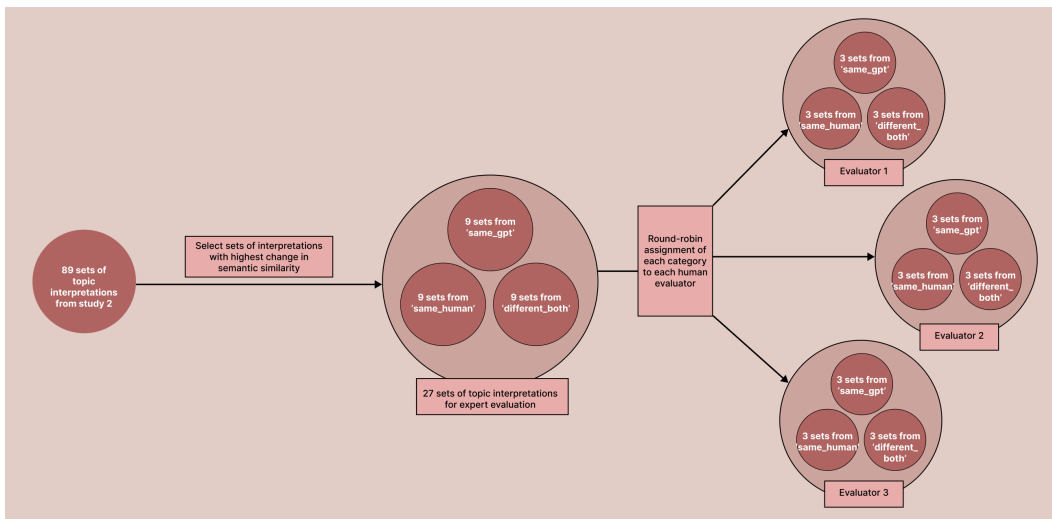


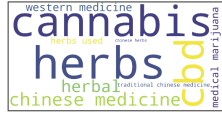
Figure 6.1: Survey construction for 3 human evaluators using the interpretations from study 2.

¹<https://www.limesurvey.org/>

6.1. Study Procedure

* Please review the topic information extracted from Colorectal Cancer patient forums:

A. Topic Wordcloud:



342 out of 290256 posts belong to this topic.

B. Topic posts:

- 23 is still posting to his blog but he did leave the forum. I miss the sometimes heated discussions between . . . Buckwith and others. I copied his blog post: I will no longer be posting to this forum. Tue, 10/08/2013 - 9:25pm — 23 The "last straw" were the posts here: The reference to TCM as being barbaric and archaic in the use of "human body parts" and other rare ingredients is about as ridiculous as it gets. If such things were used thousands of years ago, it has little if nothing to do with today's TCM treatments. Early western medicine's treatments that included "blood letting" (where a substantial amount of blood was drawn out of one's body to rid one of a disease) was as barbaric and archaic as anything else.... To this day, body parts are arbitrarily removed to western medicine's standards, to thwart a disease, even if that disease is present or not. For disease, the western medicine resolve is to remove the colon. There have been no TCM prescriptions that I have had that included anything but herbs. Nothing barbaric or archaic, and everything as natural as one can imagine. No "animal horns" or body parts: No human body parts or fecal matter, as was suggested in the link provided. (TCM) has resolved my problems with atrial fibrillation, kidney stones, blood clots, and more. I had been diagnosed with stage four colon cancer in 2006. I had been advised that only chemotherapy may provide me with time.... My surgical wounds took too long to heal (infections, etc), and they would not provide chemotherapy until the wounds were healed. Through research, I found herbs that were known to resolve problems with cancer (the list is on the blog). I never did chemo. No, it is not "proof" of anything. I may be alive and " " to this day, but it proves nothing. The fact that millions of others have used the herbs and resolved their problem with cancer is also not considered "proof" of anything.... Or not according to those that have not taken the time to explore TCM in any serious manner. Information is provided on my "blog" here. The herbs I used are listed with instructions on the same "blog". Links to TCM information are listed in that same "blog". I will no longer post here, but I am leaving behind all the information you need to consider TCM. I wish you all good health regardless of your choice of resolve....
- Is anyone on It? I am and it has been extremely helpful for me. I am curious if I am the only one and if there are people using it I would be interested in knowing if it is helping you through side effects. It has been a god send for me personally. Anyway if it is legal in your area I would encourage you to check it out, at least investigate. It may offer you great relief as it has me. I hope this is not a taboo topic but having originally scoffed at the idea of trying it I have come around to being very grateful it is an option for me. Funny story... I did not ask for permission to do this but as soon as my sister and I went to our next oncologist visit she blurted out, out of nowhere "she is on the marijuana!" I cracked up horrified but dr said that is fine how is it working and it has never been discussed since. is not treatment and I would never substitute it for chemo but it helps me so much with the side effects it is worth it to me to use it SK
- Medical it is not legal in my country but that does not make any difference. The first baby steps are being taken to make it legal. fantastic stuff. seed oil has recently been classified as a food source here and in many countries so can be bought in health food shops. that is as close as you can get to mj for many people but does the trick. it seems it is the CBD in mj that has the health benefits. is basically mj with no which is the bit that makes you stoned. some medical mj is a strain of that has very low to next to none so you will not get stoned. there are different strains for different effects. some strains for pain or appetite have higher in them so the stone can lift you above the pain or stimulate appetite. the most famous strain of medical mj is Charlottes Web this is known for it is low content. if you google 'best strains for cancer' there is much more info on the strains. I totally agree that will not cure cancer but believe it is amazing stuff that can contribute highly to keeping you well and believe we are only beginning to learn the benefits. the potential for good is huge. loved the comment "she is on the marijuana" . made me laugh so much. but i also think if you are taking medical mj or anything else please tell your doctor...

On a scale of 1 to 7, what is the **relevance** of each description to the topic.

	1	2	3	4	5	6	7
Several alternative treatment options are possible for colorectal cancer, like herbs, cannabis, medical marijuana, TMC or CBD. Some are traditional Chinese medicine, or alternative Western medicine. These substance can relieve pain, but also stimulate appetite. As such, it may relieve some symptoms that people with colorectal cancer experience during their treatment. However, it is not legal everywhere to use it for medical purposes.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The BERTopic discovered topic is about alternative treatments for colorectal cancer, such as herbs, cannabis, CBD, Chinese medicine, herbal remedies, and medical marijuana. It also mentions Western medicine as a comparison. The topic highlights the use of herbs and traditional Chinese medicine, as well as the potential benefits of medical marijuana, and how they may be used to relieve pain and stimulate appetite.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Marijuana, TMC or CBD is a substance that can relieve pain, but also stimulate appetite. As such, it may relieve some symptoms that people with colorectal cancer experience during their treatment. However, it is not legal everywhere to use it for medical purposes.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The BERTopic discovered topic is about alternative treatments for colorectal cancer, such as herbs, cannabis, CBD, Chinese medicine, herbal remedies, and medical marijuana. It also mentions Western medicine as a comparison. The topic highlights the use of herbs and traditional Chinese medicine, as well as the potential benefits of medical marijuana.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 6.2: Example of participant task for the given set of interpretations.

6. EXPERT EVALUATION OF HYBRID AND INDIVIDUAL INTERPRETATIONS

First and foremost, the evaluators were presented with the identical topic information that was used to generate the interpretations in the second study. This included topic keyphrases in a word cloud, the number of topic posts, and 3 relevant topic posts. Their main task was to assess the relevance of the interpretations to the given topic on a scale of 1 to 7. In order to prevent any potential influence on the evaluators' ratings, the order of the interpretations was randomized, ensuring that no bias was introduced based on the type of interpretation: human-written, AI-generated, individual or hybrid. An example of the task provided to the evaluators can be seen in Figure 6.2. In addition, they were given the option to comment on their method of judgment regarding the ratings they provided for the interpretations of each refinement category. This was collected to understand how the evaluators reasoned about their interpretation assessment.

6.2 Study Reflection

Once the three designed evaluation forms were completed by three proficient medical professionals from Erasmus MC, we collected the ratings they assigned to the individual and hybrid human and GPT-3.5 interpretations. These human experts have a specialization in 'Oncological and Gastrointestinal Surgery', meaning that they have the higher experience in understanding CRC patient experiences and treating CRC tumors.

P1 average ratings	AVG %	Increase	
initial_human	28.57142857	human_increase	+7.142857143
initial_gpt	38.0952381	gpt_increase	-9.523809524
improved_human	35.71428571		
improved_gpt	28.57142857		

(a) Average ratings Participant 1

P2 average ratings	AVG %	Increase	
initial_human	79.36507937	human_increase	-3.174603175
initial_gpt	71.42857143	gpt_increase	+9.523809524
improved_human	76.19047619		
improved_gpt	80.95238095		

(b) Average ratings Participant 2

P3 average ratings	AVG %	Increase	
initial_human	47.61904762	human_increase	+4.761904762
initial_gpt	28.57142857	gpt_increase	+2.380952381
improved_human	52.38095238		
improved_gpt	30.95238095		

(c) Average ratings Participant 3

Table 6.1: Average percentage ratings of each participant survey for original human, original GPT-3.5, improved human and improved GPT-3.5 topic interpretations

Table E.1 from Appendix E shows the exact ratings for the 27 chosen topics and their sets of interpretations. These topic interpretations were specifically chosen due to their notable shifts in semantic similarity, where there was a significant change from ‘initial_human’ to ‘improved_human’ and from ‘initial_gpt’ to ‘improved_gpt’, following the survey construction principles from Figure 6.1. Complete interpretations can be accessed by cross-referencing the agreement category and corresponding topic number outlined in Appendix D, as comprehensively detailed within the provided table (i.e., Table E.1).

The interpretation ratings were used to calculate the average percentage ratings for each participant survey in Table 6.1. This was done to measure the evaluation statistics for each participant on the individual and hybrid human and GPT-3.5 interpretations. The data in the table highlights that P1 only observed notable improvements in human interpretations, P2 only recognized significant improvements in GPT-3.5 interpretations, and P3 identified substantial progress in both human and GPT-3.5 interpretations.

Global Ratings	AVG %	Increase	
initial_human	51.85185185	human_increase	+2.91005291
initial_gpt	46.03174603	gpt_increase	+0.793650794
improved_human	54.76190476		
improved_gpt	46.82539683		

Table 6.2: Global average percentage ratings for original human, original GPT-3.5, improved human and improved GPT-3.5 topic interpretations

In summary, the global average percentage ratings were computed in Table 6.2, showing the overall performance of the individual and hybrid human and GPT-3.5 interpretations across all participant surveys. The results suggest that the hybrid intelligence approach, involving collaboration between human experts and AI, has led to improved human interpretations of patient experiences in colorectal cancer research. The ratings of ‘improved_human’ interpretations being higher than those of ‘initial_human’ interpretations support the effectiveness of the hybrid approach in enhancing the topic relevance of the interpretations. While the AI interpretations also show a slight improvement, the main impact seems to be on the human side. Moreover, the data pattern indicates that participants generally assigned higher ratings to human interpretations compared to those generated by GPT-3.5. The higher ratings for human interpretations may suggest that participants perceived their own expertise and contextual understanding to yield more accurate and reliable insights from the patient forum data.

Chapter 7

Conclusions and Future Work

Throughout this thesis, the primary objective was to bridge the gap between human and machine intelligence in the interpretation of colorectal cancer patient experiences extracted from patient web forums. To achieve this goal, the research followed a three-fold approach. Firstly, we investigated and compared the individual interpretations of human experts and GPT-3.5 regarding the experiences of colorectal cancer patients. Secondly, we explored the possibilities of combining human and AI interpretations to create hybrid interpretations. Lastly, we conducted an evaluation to assess the relevance of both individual and hybrid interpretations in the context of patient experience topics. These steps allowed us to gain insights into the effectiveness of hybrid intelligence in understanding and interpreting the colorectal cancer patient journey.

This chapter will discuss the implications and limitations of each study, drawing meaningful conclusions from the objectives that have been addressed. Additionally, it will explore potential directions for future research that can further enhance the hybrid intelligence approach in understanding and interpreting colorectal cancer patient data from online patient communities.

7.1 Discussions

In chapter 4, chapter 5 and chapter 6, the thesis covered the experimental setups for each of the three objectives and provided answers to the respective sub-questions. The outcomes of the three studies yield important implications and limitations, summarized as follows.

7.1.1 Implications

To address the first sub-question (i.e., *What are the differences between human and GPT-3.5 in their topic interpretations?*), a co-creation session was conducted with 5 medical stakeholders at Erasmus MC. During this session, they reviewed and interpreted colorectal cancer (CRC) patient experience topics derived from the Non-negative Matrix Factorization (NMF) topic modeling approach. A comparison between their interpretations and those generated by GPT-3.5 revealed distinct differences. Human experts exhibited unique interpretation

7. CONCLUSIONS AND FUTURE WORK

behaviors, such as prioritizing forum posts to understand and interpret CRC patient experiences, incorporating their CRC domain expertise into the interpretation process, and providing subjective insights on CRC patient health. Conversely, GPT-3.5 showcased its distinct capabilities in NLP to recognize CRC patient experiences from topic terms, faced language understanding challenges when interpreting diverse CRC patient forum posts, and occasionally disseminated misleading information related to CRC patient experiences. Differentiating between human and AI interpretation behaviors is a critical step in the CRC hybrid workflow, as it enables medical workers to understand their own interpretation capabilities and those of pre-trained language models like GPT-3.5 in CRC research.

For answering the second sub-question (i.e., *How do human experts reach interpretive agreements with GPT-3.5?*), a Flask web application was developed to facilitate interactive engagement between 12 medical researchers from Erasmus MC and GPT-3.5 in interpreting colorectal cancer patient experiences identified through BERTopic topic modeling. Through this collaborative process, both human and GPT-3.5 topic interpretations were refined, and original and hybrid interpretations were compared to identify areas of agreement. The findings revealed that in cases where human experts made changes only to GPT-3.5 interpretations (i.e., *same_human*), they primarily extended GPT-3.5 interpretations with patient experiences related to colorectal cancer treatment options and potential chemotherapy side-effects. Conversely, when changes were made only to human interpretations (i.e., *same_gpt*), the researchers mostly used GPT-3.5's interpretation as a reference to enhance their understanding of the same patient context as for *same_human*: colorectal cancer treatment options and relationship between chemotherapy and its side-effects. Additionally, when both human and GPT-3.5 interpretations were refined (i.e. *different_both*), specific areas, such as regular treatment options, alternative treatment options, and distraction techniques for colorectal cancer patients, often demonstrated significant refinements for reaching interpretive agreements. In conclusion, this study provides a novel collaborative approach for medical workers on how to effectively integrate their domain knowledge with the pre-trained language generation of GPT-3.5 for developing hybrid interpretations on colorectal cancer patient experiences extracted from patient web forums.

To address the third and final sub-question (i.e., *What is the added value of combining human and AI topic interpretations in CRC research?*), a subset of the topic interpretations from the second study were chosen to be rated by 3 human evaluators (i.e., surgeons with higher expertise in CRC healthcare) based on how well the interpretations describe the relevance of the CRC patient experience topic. Three LimeSurvey evaluation forms were constructed by picking topics that had interpretations with significant changes in semantic similarity between their original and refined versions to allow for notable differences between individual and hybrid interpretations during the evaluation process. Based on the collected survey results from the human evaluators, the added value of combining human expertise and GPT-3.5 is that human refinements have led to significant improvements in human interpretations, whereas GPT-3.5 refinements have led to slight improvements in AI interpretations. This implies that, according to the expert judgements, the hybrid intelligence approach from the second study has successfully led to more relevant human and GPT-3.5 interpretations on CRC patient forum data. Additionally, the preference for human interpretations over the ones from GPT-3.5 is evident, given that evaluators poten-

tially favored interpretations incorporating contextual understanding and domain expertise regarding CRC patient experiences. These aspects had already been observed in the initial study's observations as detailed in Table 4.2.

7.1.2 Limitations

For the first study, the GPT-3.5 interpretations of the NMF baseline were only prompt-engineered on the topic terms and not on the documents of the topic model, potentially affecting the AI's ability to fully understand the context of the patient experiences. Additionally, time constraints led to a limited number of topics being interpreted by the participants. The manual review and interpretation of patient experience topics on printed topic cards also introduced time and labor intensity.

For the second study, participants reported that the topic representation in the application was not always optimal for understanding the patient experience, and navigation issues affected their interpretation performance. These usability challenges might have impacted the quality and efficiency of the hybrid interpretation process, potentially influencing the outcomes of the study.

The third study faced constraints due to limited time and resources. Only a subset of topics was chosen for evaluation by a small group of human experts, which may have limited the breadth of insights gained from the evaluation process. Additionally, by having each selected interpretation not be rated by more than one expert (i.e., subjective rating), biases or inconsistencies could have been introduced in the evaluation, affecting the reliability of the results and discussions drawn from the study.

7.2 Conclusion

In conclusion, Figure 7.1 shows how this thesis explores the patient forum data from Cancer Survivors Network USA with topic modeling and GPT-3.5 language generation, as well as how it compares, combines and validates the CRC patient experience topic interpretations from human experts and GPT-3.5 to answer the main research question (i.e., *How can topic modeling, GPT-3.5 language generation and human expertise be combined to explore the interpretation of patient web forums in colorectal cancer (CRC) research?*). The patient web forums were first only explored with machine intelligence: LDA and NMF topic modeling, as well as GPT-3.5 language generation as a machine-driven form of topic interpretation. Subsequently, in the first study, the NMF generated topics were also interpreted by human experts and those human interpretations were then compared with the respective GPT-3.5 interpretations to identify meaningful differences between the two parties. As the research progressed, BERTopic results were also employed to explore the patient web forums, alongside their corresponding GPT-3.5 interpretations. In the second study, this combination allowed for more complex topics to be interpreted by human researchers, leading to interactive engagement between them and GPT-3.5. This collaboration aimed to refine both human and AI interpretations, resulting in the development of hybrid interpretations. Finally, a group of trained human experts validated a subset of the original human and AI interpretations together with their hybrid versions to confirm that combining human and

7. CONCLUSIONS AND FUTURE WORK

GPT-3.5 indeed leads to improved interpretations of CRC patient experience topics. All in all, this thesis highlights that hybrid intelligence for topic interpretation requires human experts and GPT-3.5 to work together by collectively interpreting patient experiences in order to reach a more comprehensive understanding for improving patient outcomes in colorectal cancer research.

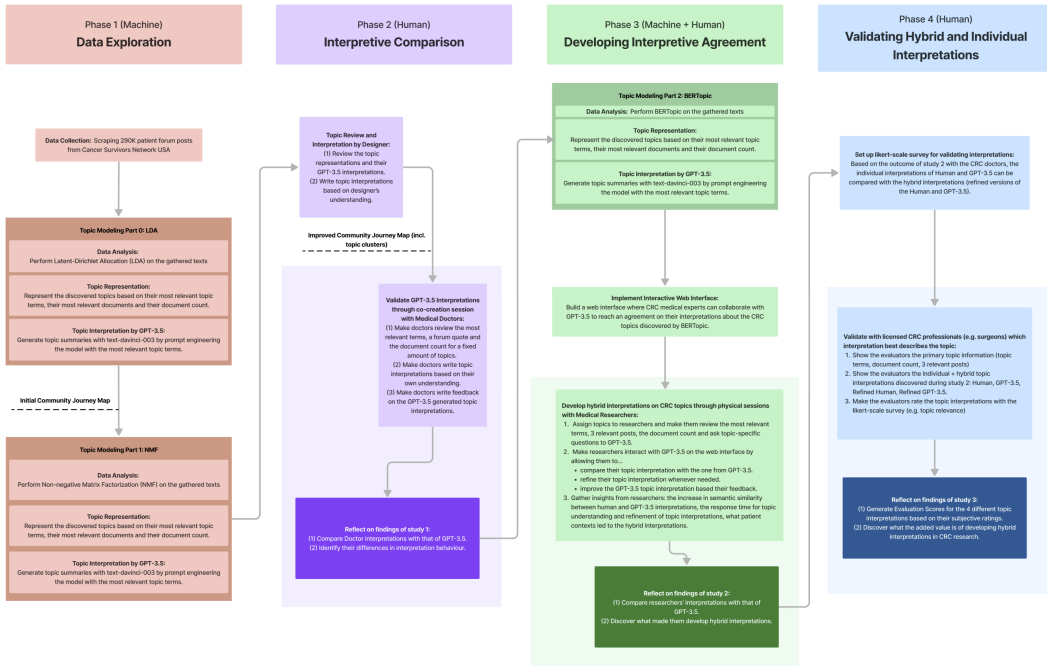


Figure 7.1: Hybrid workflow overview: step-by-step analysis for comparing, combining and validating Human and AI Interpretations in understanding CRC patient experiences.

7.3 Future work

A potential future direction for this thesis work is to engage patients in the interpretation process of colorectal cancer forum posts. Involving patients in this way ensures that the interpretations derived from online communities are truly recognizable and representative of their experiences. By including patients in the validation and refinement of interpretations made by human experts and AI models, the research can yield more meaningful and patient-centered topic interpretations of their colorectal cancer experiences.

Another notable direction involves exploring newer and more advanced large language models to assist medical researchers in interpreting patient experiences in the hybrid workflow. For instance, leveraging GPT-4 as the successor of GPT-3.5 offers more features, such as interpreting some of the top patient forum posts together with topic keywords or keyphrases due to its longer prompt size. This advancement may lead to better integration with human expertise and may further improve the hybrid interpretations developed in this thesis.

A third impactful direction is the integration of the interactive web application from the second study into clinical practice. This would streamline the process of reviewing and analyzing large volumes of patient forum data, reducing the workload for healthcare providers to understand patient experiences and provide patient-centered care. By incorporating natural language processing features of AI models like GPT-3.5, remote patient monitoring systems can be developed to offer care and support to colorectal cancer patients through digital medical devices. This integration will enable a more efficient and patient-centered approach in healthcare delivery.

Bibliography

- [1] Abeer Abuzayed and Hend Al-Khalifa. Bert for arabic topic modeling: An experimental study on bertopic technique. *Procedia Computer Science*, 189:191–194, 2021. ISSN 1877-0509. doi: <https://doi.org/10.1016/j.procs.2021.05.096>. URL <https://www.sciencedirect.com/science/article/pii/S1877050921012199>. AI in Computational Linguistics.
- [2] Amritanshu Agrawal, Wei Fu, and Tim Menzies. What is wrong with topic modeling? (and how to fix it using search-based se). *Information and Software Technology*, 98, 02 2018. doi: 10.1016/j.infsof.2018.02.005.
- [3] Zeynep Akata, Dan Balliet, Maarten de Rijke, Frank Dignum, Virginia Dignum, Gusztı Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, Hayley Hung, Catholijn Jonker, Christof Monz, Mark Neerincx, Frans Oliehoek, Henry Prakken, Stefan Schlobach, Linda van der Gaag, Frank van Harmelen, Herke van Hoof, Birna van Riemsdijk, Aimee van Wynsberghe, Rineke Verbrugge, Bart Verheij, Piek Vossen, and Max Welling. A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(8):18–28, 2020. doi: 10.1109/MC.2020.2996587.
- [4] Saleema Amershi, Maya Cakmak, W. Knox, and Todd Kulesza. Power to the people: The role of humans in interactive machine learning. *Ai Magazine*, 35:105–120, 12 2014. doi: 10.1609/aimag.v35i4.2513.
- [5] John Ayers, Adam Poliak, Mark Dredze, Eric Leas, Zechariah Zhu, Jessica Kelley, Dennis Faix, Aaron Goodman, Christopher Longhurst, Michael Hogarth, and David Smith. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine*, 04 2023. doi: 10.1001/jamainternmed.2023.1838.
- [6] Ying Bao, Xusen Cheng, Triparna de Vreede, and Gert-Jan de Vreede. Investigating the relationship between ai and trust in human-ai collaboration. 01 2021. doi: 10.24251/HICSS.2021.074.

BIBLIOGRAPHY

- [7] Kathleen Beusterien, Sarah Tsay, Shadi Gholizadeh, and Yun Su. Real-world experience with colorectal cancer chemotherapies: Patient web forum analysis. *Ecancer-medicalscience*, 7:361, 10 2013. doi: 10.3332/ecancer.2013.361.
- [8] David M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, apr 2012. ISSN 0001-0782. doi: 10.1145/2133806.2133826. URL <https://doi.org/10.1145/2133806.2133826>.
- [9] David M Blei and John D Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35, 2007.
- [10] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003. doi: 10.1162/jmlr.2003.3.4-5.993.
- [11] Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of the Biennial GSCL Conference 2009*, 01 2009.
- [12] Lucas Andreas Braun, Bijan Zomorodbakhsch, Christian Keinki, and Jutta Huebner. Information needs, communication and usage of social media by cancer patients and their relatives. *Journal of Cancer Research and Clinical Oncology*, 145:1865–1875, 2019.
- [13] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey A. Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 2018. doi: 10.3322/caac.21492.
- [14] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL <https://arxiv.org/abs/2005.14165>.
- [15] Deng Cai, Xiaofei He, Xindong Wu, and Jiawei Han. Non-negative matrix factorization on manifold. In *2008 Eighth IEEE International Conference on Data Mining*, pages 63–72. IEEE, 2008. doi: 10.1109/ICDM.2008.57.
- [16] Jaime Carbonell and Jade Stewart. The use of mmr, diversity-based reranking for reordering documents and producing summaries. *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, 06 1999. doi: 10.1145/290941.291025.
- [17] Marco Cascella, Jonathan Montomoli, Valentina Bellini, and Bignami Elena. Evaluating the feasibility of chatgpt in healthcare: An analysis of multiple clinical and

- research scenarios. *Journal of Medical Systems*, 47, 03 2023. doi: 10.1007/s10916-023-01925-4.
- [18] Dhivya Chandrasekaran and Vijay Mago. Evolution of semantic similarity—a survey. *ACM Comput. Surv.*, 54(2), feb 2021. ISSN 0360-0300. doi: 10.1145/3440755. URL <https://doi.org/10.1145/3440755>.
- [19] Anthony Chang, Tatiana Moreno, William Feaster, and Louis Ehwerhemuepha. *Towards Artificial and Human Intelligence in Hybrid Healthcare*, pages 7–16. Springer International Publishing, Cham, 2022. ISBN 978-3-031-04836-4. doi: 10.1007/978-3-031-04836-4_2. URL https://doi.org/10.1007/978-3-031-04836-4_2.
- [20] Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David Blei. Reading tea leaves: How humans interpret topic models. volume 32, pages 288–296, 01 2009.
- [21] Junhan Hannah Chen and Yuan Wang. Social media usage for health purposes: Systematic review (preprint). *Journal of Medical Internet Research*, 23, 01 2020. doi: 10.2196/17917.
- [22] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. All that’s ‘human’ is not gold: Evaluating human evaluation of generated text, 2021.
- [23] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990. doi: [https://doi.org/10.1002/\(SICI\)1097-4571\(199009\)41:6<391::AID-ASII>3.0.CO;2-9](https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASII>3.0.CO;2-9). URL <https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-4571%28199009%2941%3A6%3C391%3A%3AAID-ASII%3E3.0.CO%3B2-9>.
- [24] Dominik Dellermann, Philipp Ebel, Matthias Söllner, and Jan Marco Leimeister. Hybrid intelligence. *Business Information Systems Engineering*, 10 2019. doi: 10.1007/s12599-019-00595-2.
- [25] Dominik Dellermann, Adrian Calma, Nikolaus Lipusch, Thorsten Weber, Sascha Weigel, and Philipp Ebel. The future of human-ai collaboration: a taxonomy of design knowledge for hybrid intelligence systems, 2021.
- [26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL <http://arxiv.org/abs/1810.04805>.
- [27] David Fogel. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. *Contemporary Clinical Trials Communications*, 11, 08 2018. doi: 10.1016/j.conctc.2018.08.001.
- [28] Jeffrey Friedl. *Mastering Regular Expressions*. 08 2006. ISBN 0596528124.

- [29] Jon Fripp, Michael Fripp, and Deborah Fripp. *Speaking of science*. Newnes, 2000.
- [30] Mirta Galesic and Michael Bosnjak. Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, 73, 06 2009. doi: 10.1093/poq/nfp031.
- [31] Ariel Gera, Alon Halfon, Eyal Shnarch, Yotam Perlitz, Liat Ein-Dor, and Noam Slonim. Zero-shot text classification with self-training, 2022.
- [32] Miguel Grinberg. *Flask web development: developing web applications with python*. ” O’Reilly Media, Inc.”, 2018.
- [33] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure, 2022.
- [34] Aditi Gupta and Hoor Fatima. Topic modeling in healthcare: A survey study. *Neuro-Quantology*, 20:6214–6221, 09 2022. doi: 10.14704/NQ.2022.20.11.NQ66619.
- [35] Alexander Hoyle, Pranav Goel, Denis Peskov, Andrew Hian-Cheong, Jordan Boyd-Graber, and Ps Resnik. Is automated topic model evaluation broken?: The incoherence of coherence, 07 2021.
- [36] Rahib Imamguluyev. The rise of gpt-3: Implications for natural language processing and beyond. *International Journal of Research Publication and Reviews*, 4:4893–4903, 03 2023. doi: 10.55248/gengpi.2023.4.33987.
- [37] Mohammad Hossein Jarrahi, David Askay, Ali Eshraghi, and Preston Smith. Artificial intelligence and knowledge management: A partnership between human and ai. *Business Horizons*, 66(1):87–99, 2023. ISSN 0007-6813. doi: <https://doi.org/10.1016/j.bushor.2022.03.002>. URL <https://www.sciencedirect.com/science/article/pii/S0007681322000222>.
- [38] Mohd Javaid, Abid Haleem, and Ravi Pratap Singh. Chatgpt for healthcare services: An emerging stage for an innovative perspective. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 3(1):100105, 2023. ISSN 2772-4859. doi: <https://doi.org/10.1016/j.tbench.2023.100105>. URL <https://www.sciencedirect.com/science/article/pii/S2772485923000224>.
- [39] Jiwon Jung. Developing data-enabled design in the field of digital health. 2023.
- [40] Pooja Kherwa and Poonam Bansal. Topic modeling: A comprehensive review. *ICST Transactions on Scalable Information Systems*, 7:159623, 07 2018. doi: 10.4108/eai.13-7-2018.159623.
- [41] Diane Korngiebel and Sean Mooney. Considering the possibilities and pitfalls of generative pre-trained transformer 3 (gpt-3) in healthcare delivery. *npj Digital Medicine*, 4, 12 2021. doi: 10.1038/s41746-021-00464-x.

- [42] Marjolein Kremers, Tessel Zaalberg, Eva van den Ende, Marlou Beneden, Frits Holleman, Prabath Nanayakkara, and Harm Haak. Patient's perspective on improving the quality of acute medical care: determining patient reported outcomes. *BMJ Open Quality*, 8, 09 2019. doi: 10.1136/bmjopen-2019-000736.
- [43] Harsh Kumar, Ilya Musabirov, Jiakai Shi, Adele Lauzon, Kwan Kiu Choy, Ofek Gross, Dana Kulzhabayeva, and Joseph Jay Williams. Exploring the design of prompts for applying gpt-3 based chatbots: A mental wellbeing case study on mechanical turk, 2022.
- [44] Daniel Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–91, 11 1999. doi: 10.1038/44565.
- [45] Marieke Leeuwen, Olga Husson, Paola Alberti, Juan Arraras, Olivier Chinot, Anna Costantini, Anne-Sophie Darlington, Linda Dirven, Martin Eichler, Eva Hammerlid, Bernhard Holzner, Colin Johnson, Meropi Kontogianni, Trille Kjaer, Ofir Morag, Sandra Nolte, Andrew Nordin, Andrea Pace, Monica Pinto, and Lonneke Poll-Franse. Understanding the quality of life (qol) issues in survivors of cancer: Towards the development of an eortc qol cancer survivorship questionnaire. *Health and Quality of Life Outcomes*, 16, 06 2018. doi: 10.1186/s12955-018-0920-0.
- [46] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70, jan 2022. doi: 10.1109/tkde.2020.2981314. URL <https://doi.org/10.1109/tkde.2020.2981314>.
- [47] Junyi Li, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. Pretrained language models for text generation: A survey, 2021.
- [48] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. page 10, 01 2004.
- [49] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models, 2023.
- [50] Saskia Locke, Anthony Bashall, Sarah Al-Adely, John Moore, Anthony Wilson, and Gareth B. Kitchen. Natural language processing in medicine: A review. *Trends in Anaesthesia and Critical Care*, 38:4–9, 2021. ISSN 2210-8440. doi: <https://doi.org/10.1016/j.tacc.2021.02.007>. URL <https://www.sciencedirect.com/science/article/pii/S2210844021000411>.
- [51] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2, 03 2017. doi: 10.21105/joss.00205.

- [52] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.
- [53] Sara Mifrah and EL Habib Benlahmar. Topic modeling coherence: A comparative study between lda and nmf models using covid'19 corpus. *International Journal of Advanced Trends in Computer Science and Engineering*, 08 2020. doi: 10.30534/ijatcse/2020/231942020.
- [54] Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text, 2023.
- [55] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56, 08 2022. doi: 10.1007/s10462-022-10246-w.
- [56] Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. Using of jaccard coefficient for keywords similarity. 03 2013.
- [57] OpenAI. Gpt-4 technical report, 2023.
- [58] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.
- [59] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL <https://aclanthology.org/P02-1040>.
- [60] Marieke M.M. Peeters, Jurriaan Diggelen, Karel Bosch, Adelbert Bronkhorst, Mark Neerinx, Jan Maarten Schraagen, and Stephan Raaijmakers. Hybrid collective intelligence in a human–ai society. *AI SOCIETY*, 36, 03 2021. doi: 10.1007/s00146-020-01005-y.
- [61] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.
- [62] Partha Pratim Ray. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3:121–154, 2023. ISSN 2667-3452. doi: <https://doi.org/10.1016/j.iotcps.2023.04.003>. URL <https://www.sciencedirect.com/science/article/pii/S266734522300024X>.

- [63] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.
- [64] Frank Rosner, Alexander Hinneburg, Michael Röder, Martin Nettling, and Andreas Both. Evaluating topic coherence measures, 2014. URL <https://arxiv.org/abs/1403.6397>.
- [65] Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. A survey of evaluation metrics used for nlg systems, 2020.
- [66] Shipra Sharma. Clustering small and medium sized dutch enterprises using hybrid intelligence. Master’s thesis, Delft University of Technology, Delft, Netherlands, 2021.
- [67] Alison Smith, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. Closing the loop: User-centered design and evaluation of a human-in-the-loop topic modeling system. pages 293–304, 03 2018. doi: 10.1145/3172944.3172965.
- [68] Wilbert Tabone and Joost de Winter. Using chatgpt for human-computer interaction research: A primer, 03 2023.
- [69] Yee Teh, Michael Jordan, Matthew Beal, and David Blei. Sharing clusters among related groups: Hierarchical dirichlet processes. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004. URL <https://proceedings.neurips.cc/paper/2004/file/fb4ab556bc42d6f0ee0f9e24ec4d1af0-Paper.pdf>.
- [70] Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech Language*, 67:101151, 2021. ISSN 0885-2308. doi: <https://doi.org/10.1016/j.csl.2020.101151>. URL <https://www.sciencedirect.com/science/article/pii/S088523082030084X>.
- [71] Ike Vayansky and Sathish A.P. Kumar. A review of topic modeling methods. *Information Systems*, 94:101582, 2020. ISSN 0306-4379. doi: <https://doi.org/10.1016/j.is.2020.101582>. URL <https://www.sciencedirect.com/science/article/pii/S0306437920300703>.
- [72] M.K. Ward and Adam W. Meade. Dealing with careless responding in survey data: Prevention, identification, and recommended best practices. *Annual Review of Psychology*, 74(1):577–596, 2023. doi: 10.1146/annurev-psych-040422-045007. URL <https://doi.org/10.1146/annurev-psych-040422-045007>. PMID: 35973734.
- [73] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022.

BIBLIOGRAPHY

- [74] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt, 2023.
- [75] Zhirong Yang, He Zhang, Zhijian Yuan, and Erkki Oja. Kullback-leibler divergence for nonnegative matrix factorization. In Timo Honkela, Włodzisław Duch, Mark Girolami, and Samuel Kaski, editors, *Artificial Neural Networks and Machine Learning – ICANN 2011*, pages 250–257, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-21735-7.
- [76] Rohola Zandie and Mohammad H. Mahoor. Topical language generation using transformers, 2021.
- [77] Zoya, Seemab Latif, Faisal Shafait, and Rabia Latif. Analyzing lda and nmf topic models for urdu tweets via automatic labeling. *IEEE Access*, PP:1–1, 09 2021. doi: 10.1109/ACCESS.2021.3112620.

Appendix A

Patient Journey Map

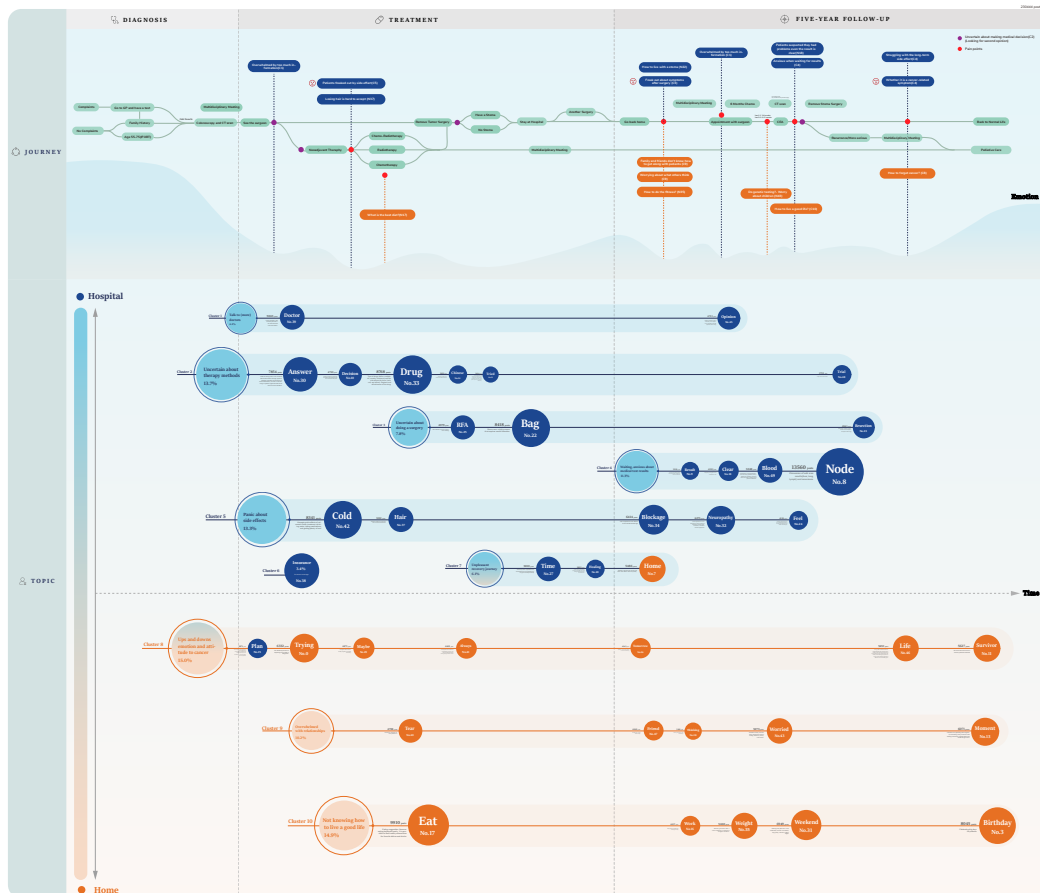


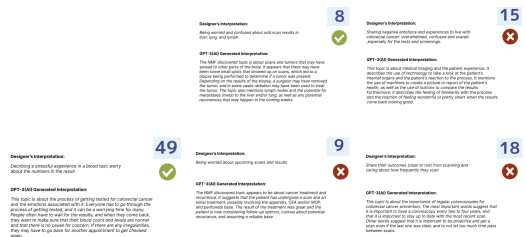
Figure A.1: Designer's Patient Journey Map after she interpreted the NMF topic modeling results for the first study

Appendix B

First Study's Individual Human and AI NMF Topic Interpretations



(a) Human

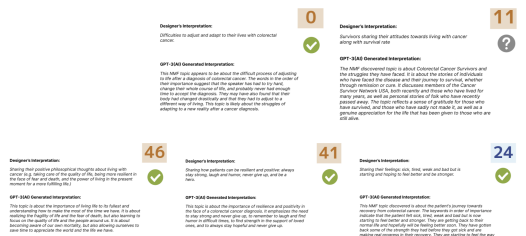


(b) AI

Figure B.1: Participant 1 - Cluster 4 Interpretations



(a) Human



(b) AI

Figure B.2: Participant 1 - Cluster 8 Interpretations

B. FIRST STUDY'S INDIVIDUAL HUMAN AND AI NMF TOPIC INTERPRETATIONS

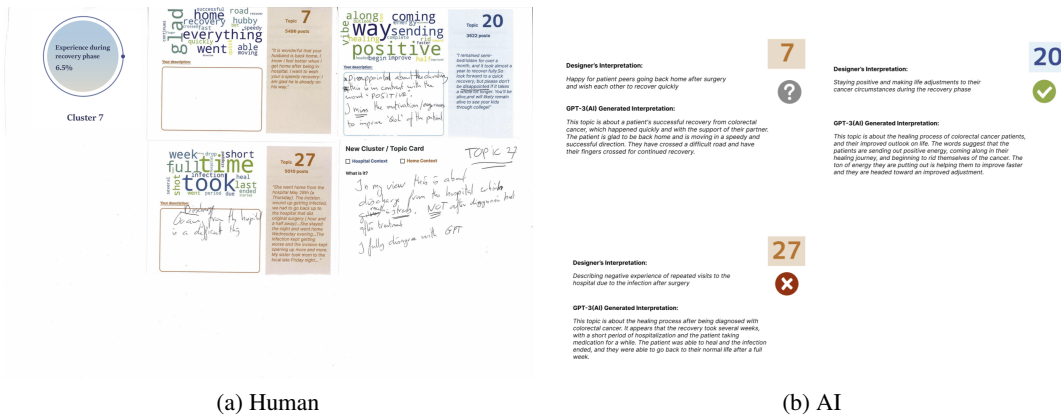


Figure B.3: Participant 2 - Cluster 7 Interpretations



Figure B.4: Participant 2 - Cluster 10 Interpretations

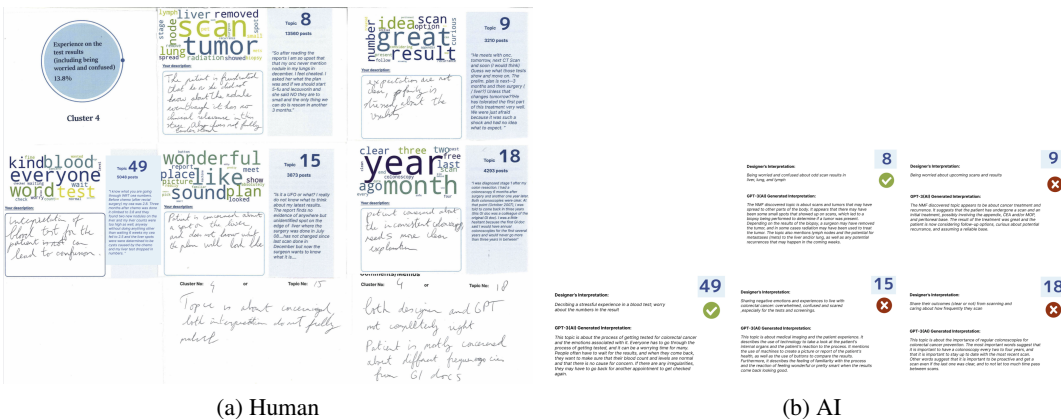
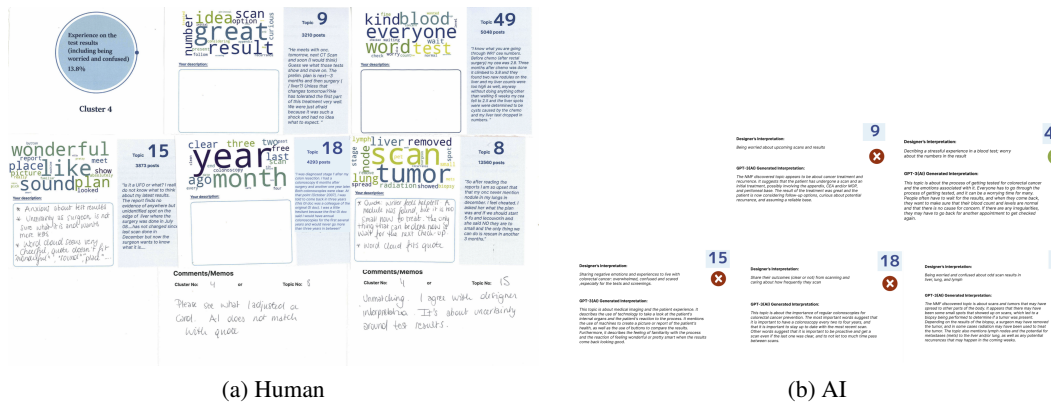


Figure B.5: Participant 3 - Cluster 4 Interpretations



(a) Human

(b) AI

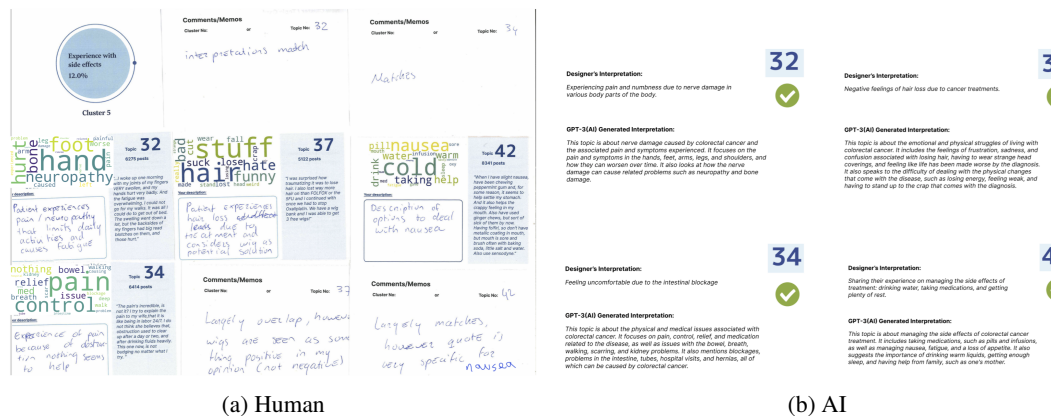
Figure B.6: Participant 4 - Cluster 4 Interpretations



(a) Human

(b) AI

Figure B.7: Participant 4 - Cluster 10 Interpretations



(a) Human

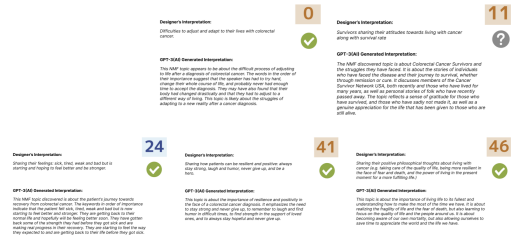
(b) AI

Figure B.8: Participant 5 - Cluster 5 Interpretations

B. FIRST STUDY'S INDIVIDUAL HUMAN AND AI NMF TOPIC INTERPRETATIONS



(a) Human



(b) AI

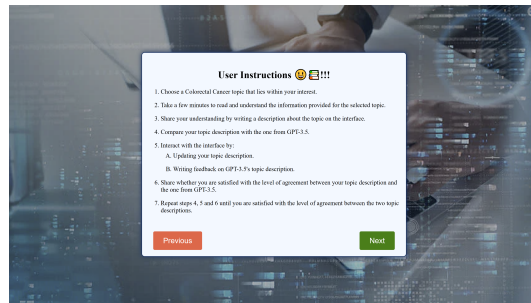
Figure B.9: Participant 5 - Cluster 8 Interpretations

Appendix C

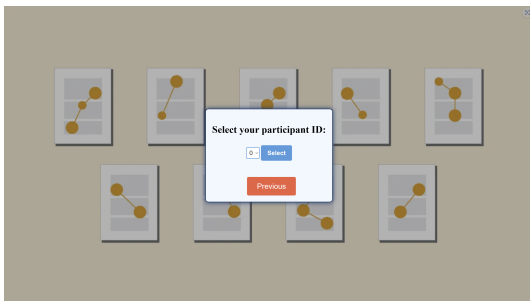
Second Study's Flask Webpages



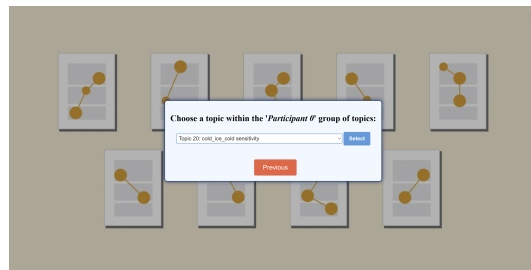
(a) Page 0



(b) Page 1



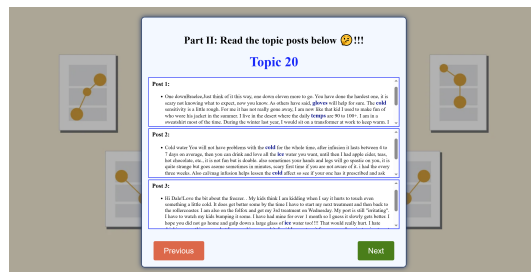
(c) Page 2a



(d) Page 2b

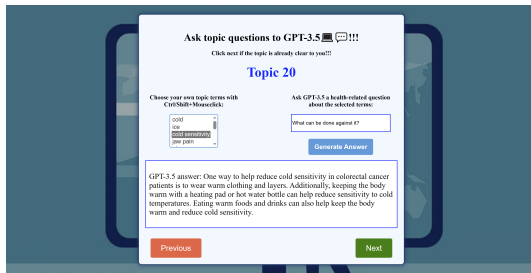


(e) Page 3



(f) Page 4

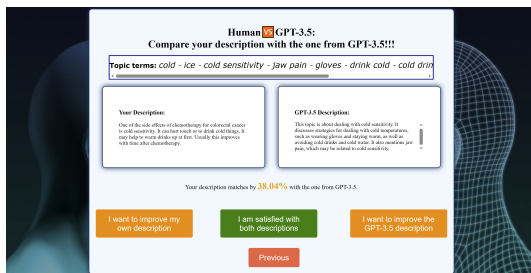
Figure C.1: Flask Web Application used to develop hybrid interpretations as part of the second study



(g) Page 5



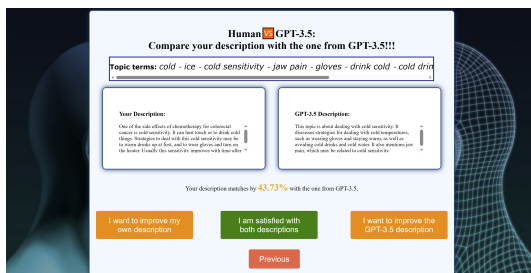
(h) Page 6



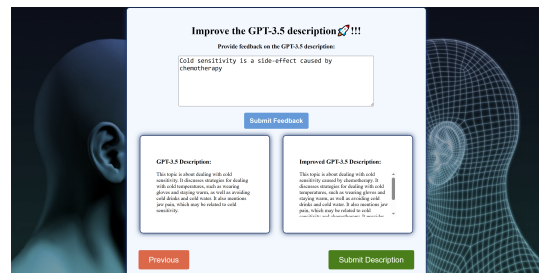
(i) Page 7



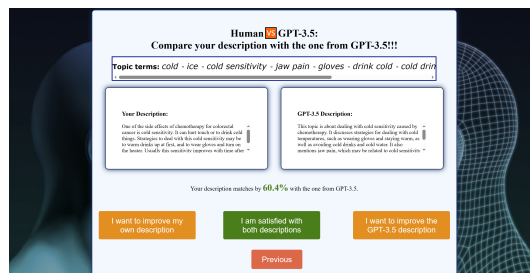
(j) Page 6 - Recurrence



(k) Page 7 - Recurrence



(l) Page 8



(m) Page 7 - Second Recurrence

Figure C.1: Flask Web Application used to develop hybrid interpretations as part of the second study (cont.)

Appendix D

Second Study's Saved Interpretations per Agreement Category

topic_id	human_description	improved_human_description	Human change	gpt_description	improved_gpt_description	AI change	Implication
72	This topic is about a patients group that calls themselves Semi-colons. It is also about a patient writing a book about their experiences.	This topic is about a patients group that calls themselves Semi-colons. It is also about a patient writing a book about their experiences.	X	This topic is about the different ways people refer to themselves when discussing their colorectal cancer experiences. It includes references to semi colons, colons, initials, middle name, nickname, semicolons, screen name, follow semi, and semi club. These terms are used to refer to the individual's identity, either in the form of a name, a nickname, or a symbol such as a semi colon.	This topic is about the different ways people refer to themselves when discussing their colorectal cancer experiences. It includes references to semi colons, colons, initials, middle name, nickname, semicolons, screen name, follow semi, and semi club. These terms are used to refer to the individual's identity, either in the form of a name, a nickname, or a symbol such as a semi colon.	X	Matching Human-AI: both describe same meaning of semi-colons (social)
14	When patients start with chemotherapy they can get neuropathy. Their feet and hands can tingle and it can be very frustrating. They are not sleeping very well. They need help to get rid of the symptoms. A solution is taking calcium or magnesium.	When patients start with chemotherapy they can get neuropathy. Their feet and hands can tingle and it can be very frustrating. They are not sleeping very well. They need help to get rid of the symptoms. A solution is taking calcium or magnesium.	X	The BERTopic discovered topic is about peripheral neuropathy, which is a type of nerve damage that causes numbness, tingling, and other sensations in the hands and feet. It can also cause worse symptoms such as pain, burning, and loss of coordination. The topic also mentions calcium and magnesium, which are important minerals for nerve health.	The BERTopic discovered topic is about peripheral neuropathy, which is a type of nerve damage that causes numbness, tingling, and other sensations in the hands and feet. It can also cause worse symptoms such as pain, burning, and loss of coordination. The topic also mentions calcium and magnesium, which are important minerals for nerve health.	X	Matching Human-AI: both describe the side-effect neuropathy (physical)
74	This topic is about dx, which is being referred to as the diagnosis of colorectal cancer. The topic consists of patients or relatives who are talking about the moment they were diagnosed and how they managed that news. It also describes a memory of being diagnosed some time ago.	This topic is about dx, which is being referred to as the diagnosis of colorectal cancer. The topic consists of patients or relatives who are talking about the moment they were diagnosed and how they managed that news. It also describes a memory of being diagnosed some time ago.	X	This topic is about the diagnosis of colorectal cancer, with particular focus on the stage of the cancer and when it was diagnosed. The keywords suggest that the topic is about recently diagnosed cases, as well as cases that were diagnosed in October and in 2009. It also mentions prior diagnosis and stage III diagnosis.	This topic is about the diagnosis of colorectal cancer, with particular focus on the stage of the cancer and when it was diagnosed. The keywords suggest that the topic is about recently diagnosed cases, as well as cases that were diagnosed in October and in 2009. It also mentions prior diagnosis and stage III diagnosis.	X	Matching Human-AI: both describe diagnosis of colorectal cancer (physical)
94	This topic is about probiotics to prevent diarrhea which can be present for chemotherapy patients. Probiotics might help.	This topic is about probiotics to prevent diarrhea which can be present for chemotherapy patients. Probiotics might help.	X	This topic is about the use of probiotics to treat digestive issues related to colorectal cancer. Probiotics are live microorganisms found in fermented foods such as yogurt, as well as in supplement form. These microorganisms can help to improve the balance of bacteria in the digestive tract, potentially reducing symptoms such as diarrhea. Other foods containing probiotics include dairy products and yeast.	This topic is about the use of probiotics to treat digestive issues related to colorectal cancer. Probiotics are live microorganisms found in fermented foods such as yogurt, as well as in supplement form. These microorganisms can help to improve the balance of bacteria in the digestive tract, potentially reducing symptoms such as diarrhea. Other foods containing probiotics include dairy products and yeast.	X	Matching Human-AI: both cover how probiotics can help prevent digestive issues like diarrhea (physical)
5	PET-CT scan for evaluation of recurrence of colorectal cancer. PET-CT scan are sensitive to uptake of glucose and therefore can be a good diagnostic test to evaluate disseminated disease or recurrence of disease. This is an advantage compared to CT scans. Tumors have a high uptake of glucose. If the uptake is slow cancer can be missed in an earlier PET.	PET-CT scan for evaluation of recurrence of colorectal cancer. PET-CT scan are sensitive to uptake of glucose and therefore can be a good diagnostic test to evaluate disseminated disease or recurrence of disease. This is an advantage compared to CT scans. Tumors have a high uptake of glucose. If the uptake is slow cancer can be missed in an earlier PET.	X	This topic is about the process of undergoing scans for colorectal cancer. It covers the anticipation of waiting for scan results, the importance of luck in obtaining good results, and the frequency of scans. It also touches on the immediacy of having a scan today.	This topic is about the process of undergoing scans for colorectal cancer. It covers the anticipation of waiting for scan results, the importance of luck in obtaining good results, and the frequency of scans. It also touches on the immediacy of having a scan today.	X	Unmatching Human-AI: Human describes scans for evaluating colorectal cancer recurrence (physical), while AI describes the process of undergoing scans (emotional)
75	This topic is about cancer survivors and the duration of survival	This topic is about cancer survivors and the duration of survival	X	This topic is about the experiences of colorectal cancer survivors, including their long-term survival, stories of survival, and the support they receive from fellow survivors, caregivers, and the Cancer Survivor Network USA board.	This topic is about the experiences of colorectal cancer survivors, including their long-term survival, stories of survival, and the support they receive from fellow survivors, caregivers, and the Cancer Survivor Network USA board.	X	Matching Human-AI: both cover CRC survivors and their experiences (social-emotional-physical)
85	This topic is about microwave radio ablation for solitary liver metastasis in colorectal cancer.	This topic is about microwave radio ablation for solitary liver metastasis in colorectal cancer.	X	This topic is about the use of ablation techniques, such as microwave and radiofrequency ablation, to treat colorectal cancer. Ablation is a minimally invasive procedure used to destroy cancerous tissue, and is often used to treat liver tumours. The weights of the keywords indicate that this topic is primarily focused on the use of ablation techniques to treat colorectal cancer.	This topic is about the use of ablation techniques, such as microwave and radiofrequency ablation, to treat colorectal cancer. Ablation is a minimally invasive procedure used to destroy cancerous tissue, and is often used to treat liver tumours. The weights of the keywords indicate that this topic is primarily focused on the use of ablation techniques to treat colorectal cancer.	X	Matching Human-AI: both cover the use of microwave and radiofrequency ablation techniques to treat colorectal cancer (physical)
95	Histology of the colorectal cancer can be divided in signet cell carcinoma which is a type of adenocarcinoma. The type of tumor can predict your prognosis. The stage of the tumor also predicts the prognosis	Histology of the colorectal cancer can be divided in signet cell carcinoma which is a type of adenocarcinoma. The type of tumor can predict your prognosis. The stage of the tumor also predicts the prognosis	X	This topic is about a type of colorectal cancer called signet ring cell carcinoma. This type of cancer is typically found in stage 3b and is characterized by a mucinous adenocarcinoma. Signet ring cells are the most common type of cells found in this type of cancer.	This topic is about a type of colorectal cancer called signet ring cell carcinoma. This type of cancer is typically found in stage 3b and is characterized by a mucinous adenocarcinoma. Signet ring cells are the most common type of cells found in this type of cancer.	X	Matching Human-AI: both cover signet ring cell carcinoma as a type of colorectal cancer (physical)
67	This topic describes an effect related to the surgery for colorectal cancer: a hernia. This is caused by the incision of the surgery, which leads to a weakening of the abdominal muscles. A common place for this to occur is the belly button. A large hernia can be repaired by a surgeon. This can be done through open surgery or laparoscopic surgery.	This topic describes an effect related to the surgery for colorectal cancer: a hernia. This is caused by the incision of the surgery, which leads to a weakening of the abdominal muscles. A common place for this to occur is the belly button. A large hernia can be repaired by a surgeon. This can be done through open surgery or laparoscopic surgery.	X	This topic is about hernia repair, which is a surgical procedure to repair weakened or torn abdominal muscles. It involves making an incision in the abdomen and repairing the hernia using laparoscopic techniques. The surgeon may also need to remove adhesions and hernias from the belly button.	This topic is about hernia repair, which is a surgical procedure to repair weakened or torn abdominal muscles. It involves making an incision in the abdomen and repairing the hernia using laparoscopic techniques. The surgeon may also need to remove adhesions and hernias from the belly button.	X	Matching Human-AI: both cover hernia related to surgery for colorectal cancer (physical)
77	This topic describes some of the side effects of chemotherapy Irinotecan, like hair loss, fatigue, nausea and severe diarrhea. Solutions can be medicine like Immodium and atropine.	This topic describes some of the side effects of chemotherapy Irinotecan, like hair loss, fatigue, nausea and severe diarrhea. Solutions can be medicine like Immodium and atropine.	X	This topic is about the side effects of Irinotecan, a chemotherapy drug used to treat colorectal cancer. The most prominent side effects discussed are hair loss, severe diarrhea, fatigue, nausea, and atropine and Immodium use to manage the diarrhea. The topic also mentions the dose of Irinotecan and the 5FU combination used in treatment.	This topic is about the side effects of Irinotecan, a chemotherapy drug used to treat colorectal cancer. The most prominent side effects discussed are hair loss, severe diarrhea, fatigue, nausea, and atropine and Immodium use to manage the diarrhea. The topic also mentions the dose of Irinotecan and the 5FU combination used in treatment.	X	Matching Human-AI: both cover side effects associated with irinotecan chemotherapy for colorectal cancer (physical)
28	This topic describes the presence of cancer in the lymph nodes and the relief patient experience if these are negative.	This topic describes the presence of cancer in the lymph nodes and the relief patient experience if these are negative.	X	This topic is about the involvement of lymph nodes in colorectal cancer staging. It covers the spread of cancer to nearby lymph nodes, as well as node involvement and lymph node involvement.	This topic is about the involvement of lymph nodes in colorectal cancer staging. It covers the spread of cancer to nearby lymph nodes, as well as node involvement and lymph node involvement.	X	Matching Human-AI: both mention the presence of cancer in lymph nodes (physical)
38	This topic describes the effect and benefits of vitamin supplements in colorectal patients. It describes the different experience patients have with said vitamins.	This topic describes the effect and benefits of vitamin supplements in colorectal patients. It describes the different experience patients have with said vitamins.	X	This topic is about the use of vitamins and supplements to manage colorectal cancer. It includes the use of vitamin D3, multivitamins, and high doses of vitamin supplements to maintain healthy vitamin levels.	This topic is about the use of vitamins and supplements to manage colorectal cancer. It includes the use of vitamin D3, multivitamins, and high doses of vitamin supplements to maintain healthy vitamin levels.	X	Matching Human-AI: both cover the experience of using vitamins and supplements (physical)
9	This topic described the port that is used to draw blood and administer chemotherapy.	This topic described the port that is used to draw blood and administer chemotherapy.	X	This topic is about the process of installing and removing ports for veins, which is a common procedure for colorectal cancer patients. It includes the steps of getting a port installed, returning blood, and luck with the port removal. It also mentions needles, which are often used in the process.	This topic is about the process of installing and removing ports for veins, which is a common procedure for colorectal cancer patients. It includes the steps of getting a port installed, returning blood, and luck with the port removal. It also mentions needles, which are often used in the process.	X	Matching Human-AI: both cover the use of ports to treat CRC (physical)
49	This topic is about genetic screening form young colorectal patients. They will be tested voo a genetic mutation, when they have a mutation like this they have a higher change for a new cancer, frequent follow-up is needed for them and their family, when they also have this mutation. follow up included colonoscopy.	This topic is about genetic screening form young colorectal patients. They will be tested voo a genetic mutation, when they have a mutation like this they have a higher change for a new cancer, frequent follow-up is needed for them and their family, when they also have this mutation. follow up included colonoscopy.	X	This topic is about genetic testing and mutations related to colorectal cancer. It covers topics such as genetic testing, mutations, screening, colonoscopies, genetics, tested positive, KRAS mutation, and genomic testing.	This topic is about genetic testing and mutations related to colorectal cancer. It covers topics such as genetic testing, mutations, screening, colonoscopies, genetics, tested positive, KRAS mutation, and genomic testing.	X	Matching Human-AI: both cover genetic testing for discovering mutations (physical)
33	This topic is about the meaning of tattoos in colorectal cancer. Tattoos can refer to the mark on the localization of the cancer to help treating it, or it can refer to a symbol or text that gave strength to a patient with colorectal cancer. Bracelets and ribbons are used to show support to colorectal cancer patients.	This topic is about the meaning of tattoos in colorectal cancer. Tattoos can refer to the mark on the localization of the cancer to help treating it, or it can refer to a symbol or text that gave strength to a patient with colorectal cancer. Bracelets and ribbons are used to show support to colorectal cancer patients.	X	This topic appears to be about wearing items to show support for colorectal cancer, such as bracelets, ribbons, and tattoos. The weights suggest that bracelets and tattoos are the most popular items, followed by ribbons and yellow fans. The words wear and tattooed also suggest that people are wearing these items to show their support.	This topic appears to be about wearing items to show support for colorectal cancer, such as bracelets, ribbons, and tattoos. The weights suggest that bracelets and tattoos are the most popular items, followed by ribbons and yellow fans. The words wear and tattooed also suggest that people are wearing these items to show their support.	X	Matching Human-AI: both cover wearing items to show support for colorectal cancer (social-emotional-physical)

Figure D.1: Saved Interpretations for 'same_both' (No refinements)

D. SECOND STUDY'S SAVED INTERPRETATIONS PER AGREEMENT CATEGORY

Page #	Human Interpretation	AI Interpretation	Agreement Category
44	This topic describes the family history of colorectal cancer. They have the faith that this is the best way for a general patient. It also describes the need of the care of new gastrointestinal and common of each other's manifestations.	This topic is about the type of being an individual. It focuses on the speed with which colorectal cancer and other gastrointestinal and common of each other's manifestations. It highlights the need for a general patient to be aware of the signs and symptoms of this disease.	Human and AI refer to focus on same topic information, initial human and AI already covered main patient experience. Human focus is retroactive of patient's experience and common of each other's manifestations. AI refers to explaining the process of gastrointestinal.
45	This topic covers endoscopy as a screening tool for polyps or pre-invasive adenoma. It covers the emotional burden for patients, such as anxiety and fear for the fear of a colonoscopy. It also covers the need for a general patient to be aware of the signs and symptoms of this disease.	This topic is about the need of endoscopy as a screening tool for polyps or pre-invasive adenoma. It covers the emotional burden for patients, such as anxiety and fear for the fear of a colonoscopy. It also covers the need for a general patient to be aware of the signs and symptoms of this disease.	Human and AI refer to focus on same topic information, initial human and AI already covered main patient experience. Human focus is retroactive of patient's experience and common of each other's manifestations. AI refers to explaining the process of gastrointestinal.
46	This topic is about the use of radiation therapy for colorectal cancer. It covers the emotional burden for patients, such as anxiety and fear for the fear of a colonoscopy. It also covers the need for a general patient to be aware of the signs and symptoms of this disease.	This topic is about the use of radiation therapy for colorectal cancer. It covers the emotional burden for patients, such as anxiety and fear for the fear of a colonoscopy. It also covers the need for a general patient to be aware of the signs and symptoms of this disease.	Human and AI refer to focus on same topic information, initial human and AI already covered main patient experience. Human focus is retroactive of patient's experience and common of each other's manifestations. AI refers to explaining the process of gastrointestinal.
47	One of the side effects of chemotherapy for colorectal cancer is cold sensitivity. It includes more about how to cope with cold sensitivity, such as using a blanket, wearing warm clothes, and using a space heater.	This topic is about the side effects of chemotherapy for colorectal cancer. It includes more about how to cope with cold sensitivity, such as using a blanket, wearing warm clothes, and using a space heater.	Human and AI refer to focus on same topic information, initial human and AI already covered main patient experience. Human focus is retroactive of patient's experience and common of each other's manifestations. AI refers to explaining the process of gastrointestinal.
48	This topic is about the side effects of chemotherapy for colorectal cancer. It includes more about how to cope with cold sensitivity, such as using a blanket, wearing warm clothes, and using a space heater.	This topic is about the side effects of chemotherapy for colorectal cancer. It includes more about how to cope with cold sensitivity, such as using a blanket, wearing warm clothes, and using a space heater.	Human and AI refer to focus on same topic information, initial human and AI already covered main patient experience. Human focus is retroactive of patient's experience and common of each other's manifestations. AI refers to explaining the process of gastrointestinal.
49	This topic is about the side effects of chemotherapy for colorectal cancer. It includes more about how to cope with cold sensitivity, such as using a blanket, wearing warm clothes, and using a space heater.	This topic is about the side effects of chemotherapy for colorectal cancer. It includes more about how to cope with cold sensitivity, such as using a blanket, wearing warm clothes, and using a space heater.	Human and AI refer to focus on same topic information, initial human and AI already covered main patient experience. Human focus is retroactive of patient's experience and common of each other's manifestations. AI refers to explaining the process of gastrointestinal.
50	This topic is about the side effects of chemotherapy for colorectal cancer. It includes more about how to cope with cold sensitivity, such as using a blanket, wearing warm clothes, and using a space heater.	This topic is about the side effects of chemotherapy for colorectal cancer. It includes more about how to cope with cold sensitivity, such as using a blanket, wearing warm clothes, and using a space heater.	Human and AI refer to focus on same topic information, initial human and AI already covered main patient experience. Human focus is retroactive of patient's experience and common of each other's manifestations. AI refers to explaining the process of gastrointestinal.
51	This topic is about the side effects of chemotherapy for colorectal cancer. It includes more about how to cope with cold sensitivity, such as using a blanket, wearing warm clothes, and using a space heater.	This topic is about the side effects of chemotherapy for colorectal cancer. It includes more about how to cope with cold sensitivity, such as using a blanket, wearing warm clothes, and using a space heater.	Human and AI refer to focus on same topic information, initial human and AI already covered main patient experience. Human focus is retroactive of patient's experience and common of each other's manifestations. AI refers to explaining the process of gastrointestinal.
52	This topic is about the side effects of chemotherapy for colorectal cancer. It includes more about how to cope with cold sensitivity, such as using a blanket, wearing warm clothes, and using a space heater.	This topic is about the side effects of chemotherapy for colorectal cancer. It includes more about how to cope with cold sensitivity, such as using a blanket, wearing warm clothes, and using a space heater.	Human and AI refer to focus on same topic information, initial human and AI already covered main patient experience. Human focus is retroactive of patient's experience and common of each other's manifestations. AI refers to explaining the process of gastrointestinal.
53	This topic is about the side effects of chemotherapy for colorectal cancer. It includes more about how to cope with cold sensitivity, such as using a blanket, wearing warm clothes, and using a space heater.	This topic is about the side effects of chemotherapy for colorectal cancer. It includes more about how to cope with cold sensitivity, such as using a blanket, wearing warm clothes, and using a space heater.	Human and AI refer to focus on same topic information, initial human and AI already covered main patient experience. Human focus is retroactive of patient's experience and common of each other's manifestations. AI refers to explaining the process of gastrointestinal.
54	This topic is about the side effects of chemotherapy for colorectal cancer. It includes more about how to cope with cold sensitivity, such as using a blanket, wearing warm clothes, and using a space heater.	This topic is about the side effects of chemotherapy for colorectal cancer. It includes more about how to cope with cold sensitivity, such as using a blanket, wearing warm clothes, and using a space heater.	Human and AI refer to focus on same topic information, initial human and AI already covered main patient experience. Human focus is retroactive of patient's experience and common of each other's manifestations. AI refers to explaining the process of gastrointestinal.
55	This topic is about the side effects of chemotherapy for colorectal cancer. It includes more about how to cope with cold sensitivity, such as using a blanket, wearing warm clothes, and using a space heater.	This topic is about the side effects of chemotherapy for colorectal cancer. It includes more about how to cope with cold sensitivity, such as using a blanket, wearing warm clothes, and using a space heater.	Human and AI refer to focus on same topic information, initial human and AI already covered main patient experience. Human focus is retroactive of patient's experience and common of each other's manifestations. AI refers to explaining the process of gastrointestinal.

Figure D.4: Saved Interpretations for 'different_both' (Refinements made by both human and GPT-3.5)

Appendix E

Third Study's Interpretation Evaluation Results

E. THIRD STUDY'S INTERPRETATION EVALUATION RESULTS

P1 ratings					
Category\Topic nr.					
different_both	48	20	37	AVG	AVG %
initial_human	2	2	2	2	28.57142857
initial_gpt	2	2	2	2	28.57142857
improved_human	5	2	2	3	42.85714286
improved_gpt	2	2	2	2	28.57142857
same_human	32	55	25	AVG	AVG %
initial_human	2	2	2	2	28.57142857
initial_gpt	2	2	5	3	42.85714286
improved_gpt	2	2	2	2	28.57142857
same_gpt	2	79	34	AVG	AVG %
initial_human	2	2	2	2	28.57142857
initial_gpt	2	5	2	3	42.85714286
improved_human	2	2	2	2	28.57142857
P2 ratings					
Category\Topic nr.					
different_both	80	56	73	AVG	AVG %
initial_human	5	2	5	4	57.14285714
initial_gpt	5	6	4	5	71.42857143
improved_human	5	2	6	4.333333333	61.9047619
improved_gpt	7	6	5	6	85.71428571
same_human	62	70	22	AVG	AVG %
initial_human	6	6	6	6	85.71428571
initial_gpt	3	4	2	3	42.85714286
improved_gpt	6	3	7	5.333333333	76.19047619
same_gpt	13	86	96	AVG	AVG %
initial_human	6	7	7	6.666666667	95.23809524
initial_gpt	7	7	7	7	100
improved_human	6	6	7	6.333333333	90.47619048
P3 ratings					
Category\Topic nr.					
different_both	87	60	45	AVG	AVG %
initial_human	2	7	7	5.333333333	76.19047619
initial_gpt	4	3	2	3	42.85714286
improved_human	4	1	2	2.333333333	33.33333333
improved_gpt	3	3	2	2.666666667	38.0952381
same_human	51	52	58	AVG	AVG %
initial_human	1	2	5	2.666666667	38.0952381
initial_gpt	1	1	2	1.333333333	19.04761905
improved_gpt	1	3	1	1.666666667	23.80952381
same_gpt	50	43	92	AVG	AVG %
initial_human	1	1	4	2	28.57142857
initial_gpt	3	1	1	1.666666667	23.80952381
improved_human	4	6	5	5	71.42857143

Table E.1: Expert Evaluation Results: Interpretation ratings on a scale of 1 to 7 from professional Human Evaluators at Erasmus MC