Original Paper

Topic-Aware Summarization of Lived Health Care Experiences: Large Language Model Evaluation Study

Maneesh Bilalpur[1], MS; Megan E Hamm[2], PhD; Young Ji Lee[3,4], PhD, RN; Natasha G Norman[2], MSW; Kathleen M Mctigue[2], MD; Yanshan Wang[1,4,5], PhD

[1]Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, United States
[2]Department of Medicine, University of Pittsburgh, Pittsburgh, PA, United States
[3]School of Nursing, University of Pittsburgh, Pittsburgh, PA, United States
[4]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, United States
[5]Department of Health Information Management, University of Pittsburgh, Pittsburgh, PA, United States

Corresponding Author:

Yanshan Wang, PhD
Intelligent Systems Program
University of Pittsburgh
4200 Fifth Avenue
Pittsburgh, PA 15260
United States
Phone: 1 4123832712
Email: yanshan.wang@pitt.edu

Abstract

Background: Existing work to understand adults' health care experiences has focused on the analysis of patient feedback provided as written responses to after-visit surveys or social media discourse. Often, such written feedback has been studied using natural language processing techniques, such as topic detection and sentiment analysis, to provide coarse-grained insights. Storytelling is a powerful form of communication and may provide insights into factors contributing to gaps in health care outcomes and avenues for improvement. In addition, studying health care experiences using natural language processing techniques has been limited to patients. The experiences of stakeholders, such as caregivers and health care providers, remain underexplored.

Objective: We extract fine-grained insights from health care experiences through narratives collected from patients, caregivers, and health care providers using large language models (LLMs). Topic detection, together with hierarchical summarization of long-form stories from individuals, offers fine-grained insights. Furthermore, the study demonstrates that generated summaries can be evaluated using the LLM-as-a-judge framework and validates the outcomes through comparisons with 2 domain experts.

Methods: Fifty automatically transcribed stories of African American experiences were used to identify topics in their experiences using the latent Dirichlet allocation (LDA) technique. Stories about a given topic were summarized using an open-source LLM-based hierarchical summarization approach. Topic summaries were generated by summarizing across story summaries for each story that addressed a given topic. The generated topic summaries were rated for fabrication, accuracy, comprehensiveness, and usefulness by the GPT-4 model; its reliability was validated against the original story summaries by 2 domain experts.

Results: Whisper-based automatic transcription of audio narrations achieved a Levenshtein score of 6%. Twenty-six topics were identified using LDA and labeled using the LLM in the 50 African American stories. The GPT-4 ratings suggest that topic summaries were free from fabrication, highly accurate, comprehensive, and useful. The reliability of GPT ratings compared to expert assessments showed moderate-to-high agreement (Bennett *S*-score of 0.65 or higher). Our approach identified African American experience-relevant topics, such as health behaviors, interactions with medical team members, caregiving, and symptom management, among others. Such insights could help researchers learn from unstructured datasets in an efficient manner—leveraging the communicative power of storytelling.

Conclusions: The use of LDA and LLMs to identify and summarize the experiences of African American individuals suggests a variety of possible avenues for health research and possible clinical improvements to support patients and caregivers, thereby improving health outcomes.

JMIR Med Inform 2026;14:e85960; doi: 10.2196/85960

Keywords: topic modeling; text summarization; health disparities; unstructured qualitative data; large language models; natural language processing

Introduction

Motivation

Storytelling is a narrative style that captures information beyond objective facts by catering to the beliefs and emotions of both the narrator and the audience. Sharing lived experiences with others through stories can overcome the limitations caused by extreme objectivity in alternatives such as didactic communication [1,2]. Storytelling can successfully share health knowledge [1-4], potentially providing insight into how illness shapes people's lives, promoting behavior change, and suggesting avenues for improving health care [5-11]. Such stories—sharing the topics deemed most important by the storyteller—can suggest innovative avenues for interventions to improve the health care ecosystem. Hence, there is a need for the health informatics community to investigate narratives of health care experiences.

Despite the efficacy of narratives in health care, their usage poses certain challenges. Information stored as oral recordings is less accessible and searchable than structured data. Furthermore, moving from a single anecdote to the derivation of common themes can be expensive and time-consuming. The use of natural language processing (NLP) techniques may help to overcome these challenges. Prior studies have examined structured feedback from patients [12], free-text feedback from social media [13], and electronic health records [14] to identify themes or topics using NLP techniques. Often, topic detection has been paired with sentiment analysis to infer an individual's sentiments toward the topic. Such solutions offer automated approaches for understanding limitations in health care environments, the efficacy of treatment approaches, and overall experiences in patient care to provide tailored solutions [12]. However, existing literature has largely focused on structured, written-language data, with relatively limited exploration of unstructured sources such as spoken-dialogue narratives.

Spoken-dialogue narratives present a unique set of challenges; they are syntactically and semantically different from written language [15], are often lengthy, and tend to cover multiple topics within a single dialogue. Under such conditions, topic-specific sentiment analysis becomes challenging due to the lack of defined topic boundaries in the data. Summarization approaches address this limitation and are also more informative, as they provide fine-grained analysis compared to single sentiment labels derived from conventional sentiment analysis. The recent rapid expansion in applications of LLMs has demonstrated that they excel at summarizing documents across domains (including health care [16]) and have shown their efficacy in understanding spoken language [17]. Thus, they offer an effective solution for summarizing spoken dialogue. Despite their ability to process complex patterns, LLMs are constrained by their input context length. For example, the popular open-source LLM, Llama 3.2, has an input context length of 128K tokens. However, topics tend to cover multiple stories whose lengths often exceed the fixed context length of LLMs. In addition, long-form inputs (commonly encountered in spoken-dialogue narrations) pose a challenge, as they run the risk of forgetting intermediate portions of the document [18]. On the other hand, classical methods are not susceptible to such limitations, especially during inference. In this work, we evaluate the capabilities of LLMs in long-form dialogue understanding by leveraging the advantages of classical methods in topic detection to build a topic-aware summarization model for understanding spoken-dialogue narratives about health care experiences.

In this work, spoken narratives of experiences (referred to as "stories") from the African American population are analyzed to identify underlying topics and summarize issues raised about the storytellers' lived health care experiences. The stories are a subset of those stored in the MyPaTH Story Booth archive (termed Story Booth). We focus on stories by African American storytellers because, compared to White individuals in the United States, African Americans experience worse health outcomes and are less likely to receive health care services [19-26], despite numerous efforts to close such gaps [26-28]. Additionally, storytelling may be a particularly appealing avenue for understanding African Americans' health experiences since the community has a strong oral tradition [4,9,10] and may also foster trust-building [24].

We focus on using NLP techniques on stories of health care experiences to identify possible factors contributing to gaps in health care delivery systems, as well as potential avenues for intervention. Storytellers include patients, caregivers, and health care professionals. Our solution offers insights through fine-grained summaries per topic, compared to existing works that are limited to coarse sentiment outcomes. Topic detection was performed over transcriptions of the long-form spoken-dialogue dataset using latent Dirichlet analysis. The detected topics were augmented with topic labels for clinical interpretability using LLMs. Furthermore, we introduce topic-aware summarization, a hierarchical approach to generate topic summaries through summaries of individual stories by topic (termed topic story summary), extracting the nature of the health care experience by each topic from stories, as depicted in the topic story summaries, using LLM-based summarization.

Background

Attempts to develop automated solutions for topic detection and sentiment analysis often involve manual annotations. One such approach [29] found appointment access and wait times, empathy, explanation, friendliness, practice environment, and overall experiences as frequent topics in patient feedback. Similarly, another study [14] analyzed electronic health records to identify goals of care in patient care. Topic detection was used to identify suicide profiles from 300,000 decedents [30]. Researchers found that suicide profiles broadly covered 5 classes: mental health and substance problems, mental health problems, crisis, alcohol-related, and intimate partner problems, physical health problems, and polysubstance problems. Furthermore, they found demographic shifts in the suicide profiles with time showing an evolving landscape of health care needs. A recent review [12] of NLP techniques to understand patient-experience feedback compared literature in terms of machine learning techniques (supervised vs unsupervised) and data collection approaches (social media vs structured surveys). Among supervised approaches, the Naïve Bayes classifier was the best performing, while the unsupervised latent Dirichlet allocation (LDA) model was the popular alternative for topic detection. In addition to sentiment detection from human-annotated labels of patient comments, affordable annotation approaches were recently explored [31]. Using LLMs for annotating patient comments revealed that despite their ability to detect in zero-shot and few-shot settings, LLMs underperformed compared to human annotators.

A recent survey [32] confirmed that most existing literature on understanding patient experience is limited to topic detection and sentiment analysis of written feedback. The ease of data collection through the structured written language has led to its popularity in understanding health care experience feedback. Despite its advantages, the use of narrative-based approaches has been limited [32-34]. One study examined supervised models for topic detection in patient-doctor conversations [35]. The researchers used turn-level manual annotations as a gold-standard reference. They varied the input context length for the supervised models and found that topic detection improves with an increase in input context length. Summaries of spoken-dialogue medical conversations between patients and doctors were studied [36] for commercial applications such as note-taking. This line of work, which focuses on summarizing patient-doctor conversations, has been gaining attention in NLP and informatics communities [36,37]. We build upon this work by shifting the focus to the health care experiences of adults, shared through storytelling. Our narratives are focused on the quality of clinical interactions, their outcomes, and the broader impact on the day-to-day lives of patients, providers, and caregivers to identify potential causes and areas of health care disparity.

Our work leverages fine-grained details in participants' conversational style data about their health care experiences. We believe that such details offer a nuanced understanding of problems in the health care experiences of marginalized communities. This is achieved by using a topic-aware hierarchical summarization approach to health care feedback. In the following sections, we describe the dataset, techniques for topic detection, and hierarchical summarization, and present the findings about health care experiences from the generated summaries.

Methods

Dataset

The MyPaTH Story Booth [38] archive is a collection of individual experiences, told from the perspective of a patient, caregiver, or health care provider. Developed as community engagement infrastructure for the PaTH Clinical Research Network, the archive includes over 1500 stories related to experiences with illness, efforts to maintain health, and interactions with health care systems [39]. Story recordings were conducted in-person or by telephone. Participants can opt to share an unstructured story or to answer prompts selected from a preapproved list. They are asked to limit their stories to 20 minutes in length or less. Funding for the MyPaTH Story Booth project was awarded in 2015, and recruitment began in March 2016. Between March 24, 2016, and August 8, 2024, 1266 stories were collected by the University of Pittsburgh Story Booth team; 167 stories were provided by African American participants (n=167, 13.2%). As of February 6, 2026, 1537 stories had been recorded in the archive.

The dataset used in this work is limited to 1120 stories—the full set collected by the University of Pittsburgh staff and fully processed at the outset of this study. From this set, 50 stories from African American participants were randomly selected for summarization and analysis. OpenAI's Whisper model [40] was used for diarization (speaker detection and transcription) of the audio recordings. Due to the limited involvement of the interviewers, we focused only on the participants' statements. We validated the Whisper transcriptions against manual transcriptions from the 50 validation stories to observe 35,030 out of 565,372 character-level changes (insertions, deletions, or substitutions), that is, a Levenshtein distance [41] of 6%. Manual inspection found that the diarization quality was satisfactory and that differences in labeling conventions between manual and automatic Whisper approaches were the major source of diarization errors.

Ethical Considerations

The University of Pittsburgh's Human Research Protection Office reviewed this study (protocols STUDY19020307 and STUDY20110315) and approved it, with a determination of no greater than minimal risk.

Stories in the database are collected with informed consent and recorded as audio files. For privacy, storytellers are instructed to avoid naming people or places. Recorded stories are reviewed, and specific identifiers (eg, names and places) are redacted, but it is possible that voices could be recognized. Narratives are included on the public Story Booth

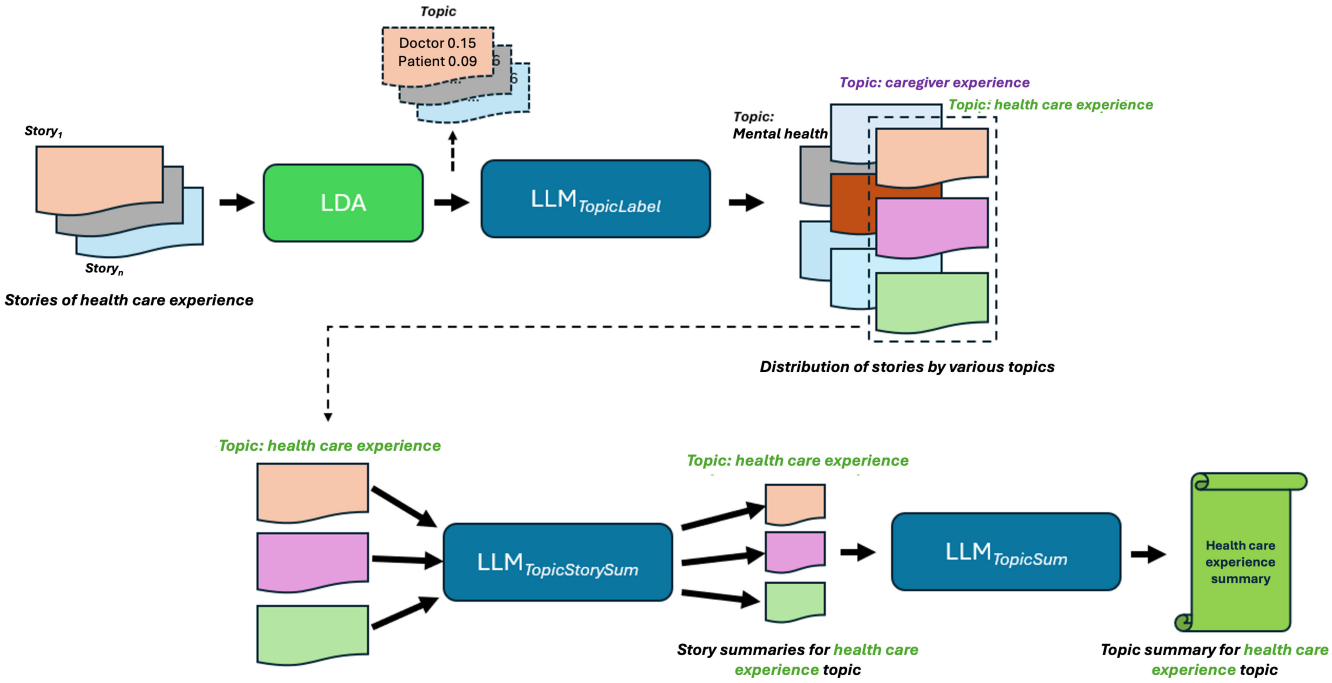
website only with the storytellers' permission. When stories are elicited to learn about the perspectives of a specific population, storytellers are compensated at the rate of US $100 per story. Compensation policies have evolved over time, and earlier stories were collected without compensation.

Analysis

Overview

The method for topic-aware summarization of long-form health care experiences combined classical methods of topic detection with recent advances in LLMs through a multistep process. First, topic detection was performed using the LDA approach [42]. The identified topics were then labeled for interpretability using LLMs, and topic story summaries were generated. Finally, topic labels and story-topic distributions from the LDA were leveraged for topic-aware summarization using a hierarchical summarization approach powered by LLMs. Individual steps are further described in the *Topic Detection and Labeling* and *Hierarchical Summarization* sections. Figure 1 shows various components of the proposed approach.

Figure 1. Various steps in the proposed approach for a topic-aware large language model (LLM) for long-form spoken-dialogue summarization. LDA: latent Dirichlet allocation.



Topic Detection and Labeling

To understand recurring topics from health care experience stories, it was necessary to identify the topics across stories. Topic modeling approaches such as LDA are often the go-to solutions for such problems and are widely used in studying health care experiences, as discussed in the related work. LDA is a probabilistic topic modeling method that captures the likelihood of topics across documents (here, stories) and simultaneously the distribution of words in each topic. LDA was used to detect topics across all 1120 stories from the dataset. The topic detection model was tuned by varying the number of topics between 50 and 1000 in steps of 50. The optimal number of topics was identified using a perplexity criterion on our validation set.

The word distributions in each topic offered a naive understanding of topics. They presented the likelihood of words in each topic but were often hard to interpret. They might also contain contradictory, ambiguous, or unrelated words under the same topic. This issue was more prevalent in the dataset, considering that the participant perspectives included patients, caregivers, and health care professionals, each offering a variety of perspectives, which made understanding the conversational themes more challenging. To overcome this limitation in interpretation, word clouds of topics were labeled using a pretrained Llama-3.1 model [43]. The Llama is an open-source LLM with competitive performance against closed-source alternatives such as GPT-4 [44] and GPT-3.5 Turbo [45]. For all LLM experiments, the 70B parameter model was used due to its long context length of 128K tokens. The inference time was optimized using the *llama.cpp* [46] framework with 4-bit quantization and GPT-Generated Unified Format encoding.

Given the challenging nature of the task, labels for each topic were derived within the context of each story. The LLM input consisted of a story and the word list from the topic-word distribution (Multimedia Appendix 1). The words were ranked in the order of their likelihood in the topic as determined by the LDA. Topic labels derived in the context of a story tend to overfit to a specific story. Moreover, across multiple stories, the topic label was found to be paraphrases of one another. To establish the consistency of topic labels

across all relevant stories, the most frequent label of the topic across all stories was chosen as the final topic label.

Hierarchical Summarization

Figure 2 presents a step-by-step overview of the approach. The LDA-identified topics were leveraged to summarize the stories from the validation set. To overcome the input context length limitations, a hierarchical summarization approach was used. The hierarchical summarization approach involved 2 steps. In the first step, topic summarization for each story was performed (Multimedia Appendix 1). This greatly consolidated the input and overcame the limitation of input context length for long-form summarization with LLMs. Following this, individual topic story summaries were further summarized to generate a holistic summary of all the stories under the topic (Multimedia Appendix 1). The hierarchical summarization also offered interpretability through tracing elements of the holistic summary.

Figure 2. Topic-aware hierarchical summarization using large language models (LLMs). LDA: latent Dirichlet allocation.

Algorithm 1 Topic-aware hierarchical summarization using Large-Language Model

Require: Dataset $D = \{s_1, s_2, ..., s_n\}$ of n stories. D_{valid} is the validation set.
Require: $LLM_{TopicLabel}, LLM_{TopicStorySum}, LLM_{TopicSum}$ be the LLM configured for topic labeling, summarizing a story given topic, and summarizing a topic given one or more story summaries respectively.
Ensure:
$TopicStory, WordTopic \leftarrow LDA(D)$ ▷ train LDA for likelihood for topic-story and word-topic distributions
$TopicStory_{valid}, WordTopic_{valid} \leftarrow LDA(D_{valid})$ ▷ predict likelihood for topic-story and word-topic distributions on D_{valid}
Let $Topics_{valid}$ be the set of topics where $TopicStory_{valid} <= TopicProbThresh$ ▷ TopicProbThresh = 0.05
for topic t in $Topics_{valid}$ **do**
  $TopicLabels \leftarrow list()$ ▷ label for topic t inferred from each story s
  for story s in D_{valid} **do**
    $TopicLabels \leftarrow LLM_{TopicLabel}(storys, WordTopic_{w,t})$
  end for
  $TopicLabel_t \leftarrow mode(TopicLabels)$ ▷ most common label across stories is used as the final topic label
  for story s in D_{valid} **do**
    $TopicStorySum_{t,s} \leftarrow LLM_{TopicStorySumm}(s, TopicLabel_t)$
  end for
  $TopicSum_t \leftarrow LLM_{TopicSum}(TopicStorySum_{t,.}, TopicLabel_t)$
end for
return $TopicSum$

Evaluation

A strong evaluation of generative models was critical to understanding their limitations. Conventional evaluation metrics, such as ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [47], only account for co-occurrence similarity between generated and reference text. Hence, in the context of LLMs, these similarity-based metrics make little sense. LLM-generated content is also prone to a distinct set of challenges, such as hallucinations and inadvertent safety limitations, which are not studied using such conventional metrics. To account for these shortcomings, we used the multidimensional evaluation framework developed with a focus on the quality of information, understanding and reasoning, expression style and persona, safety and harm, and trust and confidence principles, referred to as QUEST [48], to evaluate the generated summaries. The evaluation dimensions for this summarization task include comprehensiveness, fabrication, accuracy, and usefulness.

Human evaluation was often expensive and time-consuming, so the LLM's ability to generate topic summaries was assessed using the LLM-as-a-judge approach. To eliminate self-enhancement bias in LLMs, GPT-4 Turbo was used as the judge. GPT-4 was found to correlate well with human raters and was less susceptible to position bias [49]. Individual topic summaries, together with the respective topic story summaries, were used as input to rate the 4 QUEST dimensions on a 5-point Likert scale (see Table S1 in Multimedia Appendix 2 for the definition of our Likert scale). The reliability of the GPT responses was quantified against 2 domain experts as human raters. Raters were assigned 4 topics (chronic pain management, medical treatment, caregiving experience, and health concerns) to rate the QUEST dimensions on the same Likert scale as the GPT model. Discrepancies in Likert values between the 2 human raters were discussed and adjudicated by those raters to provide a final rating, which was compared to the GPT responses. The Bennett S-score [50] was used with quadratic weighting to calculate agreement between the raters and the GPT responses.

Results

Topic Detection and Labeling

The LDA model yielded 150 topics on the training set. For the validation set, we inferred document-level topic distributions and applied a threshold on topic probabilities to identify highly relevant topics for each story and to avoid spurious assignments. For stories where the topic probability did not exceed the threshold, the most likely topic was assigned. This approach allowed us to minimize the adverse downstream effects in the hierarchical summarization pipeline due to

incoherent topic assignment. Using this approach, 40 topics were identified in the validation set.

Topic labels for the LDA topics were necessary for the topic-aware summarization approach. Such labels also improved topic interpretability for domain experts. As mentioned in the *Methods* section, the topic labeling approach using LLM resulted in 26 topic labels across all the stories (see Table 1 for topic labels, where the numbers in parentheses indicate the number of stories that address each topic), while the remaining 14 topics were omitted due to the inherent difficulty in labeling topics based on the distribution of words and redundancy among the labels. These topic labels were then used in the summarization step.

Table 1. Topic labels derived from the large language model (LLM).

Topic labels derived from the large language model	Number of stories that address the topic
Health care experience	(38)
Healthy eating	(13)
Doctor-patient relationship	(11)
Cancer treatment	(10)
Hospital experience	(9)
Doctor experience	(6)
Personal experience	(5)
Disease diagnosis	(3)
Health care	(3)
Caregiving	(2)
Cancer diagnosis	(1)
Emotional and physical impact	(1)
Heart health	(1)
Symptoms and hospitalization	(1)
Health care system	(26)
Caregiving experience	(11)
Symptom management	(11)
Chronic pain management	(10)
Surgical experience	(7)
Medical treatment	(6)
Health concerns	(3)
Mental health	(3)
Diagnosis	(2)
Diabetes management	(1)
Health challenges	(1)
Medical condition	(1)

Hierarchical Summarization

The topic labels, together with the corresponding stories, were used to generate topic summaries through the intermediate topic story summaries. The results of hierarchical summarization for 2 sample topics are provided in Table 2 (see Table S1 in Multimedia Appendix 3 for summaries from corresponding topics). We found that the summaries for each topic were comprehensive, coherent, and aligned with the topic label derived based on the LDA topic-word distribution.

Table 2. Sample summaries for 2 different topics from our hierarchical summarization approach.

Topic label	Topic summary
Chronic pain management	Participants dealing with chronic pain management reported various challenges and experiences. Many struggled with inadequate pain management, including ineffective treatments, lack of effective pain medication, and hesitation from health care providers to treat their pain. Some participants felt that their pain was not taken seriously, leading to frustration and increased severity. Others found alternative methods, such as cannabidiol oil, yoga, and herbs, to be helpful in managing their pain. A few participants benefited from working with supportive doctors, pain clinics, and transportation programs that helped them access necessary care. Despite these efforts, many participants continued to experience chronic pain, numbness, and debilitating symptoms, affecting their daily lives and relationships.
Caregiver experience	Caregivers face numerous challenges, including lack of support, financial struggles, and emotional toll. Many participants expressed the importance of compassion, patience, and affection when caring for loved ones, particularly those with dementia or Alzheimer disease. Some participants had to navigate complex health care systems, advocate for their loved ones' needs, and deal with the emotional impact of caregiving. Others appreciated the support of family members, health care providers, and online support groups. Several participants

Topic label	Topic summary
	highlighted the need for respite care, financial assistance, and guidance for caregivers. Some also emphasized the importance of acknowledging stress and seeking help. Additionally, participants noted that health care providers should advise patients on healthy living and provide resources for caregivers. Overall, caregiving experiences varied, but many participants shared the common theme of needing more support and resources to effectively care for their loved ones.

Evaluation

We evaluated the generated summaries alongside topic story summaries to perform a human evaluation. Table 3 presents the interrater and GPT-rater agreements derived from our evaluation approach. A moderate-to-high agreement was noticed between the GPT responses and the raters. The level of agreement was particularly strong for the fabrication and comprehensiveness of the generated summaries. This suggested that the GPT evaluation of generated summaries overlapped with human raters and could be used reliably to evaluate the topic-aware hierarchical summarization model. GPT responses for all topics included in the validation set are provided in Table S2 in Multimedia Appendix 2. The accuracy of the generated summaries was consistently good for most topics, except for *diagnosis,* where the generated summary likely changed the interpretation of the topic because of one substantive point within it. The usefulness of the summaries was found to be mostly high, indicating that the approach was promising to use in place of manually identifying themes in the story and summarizing them, as defined by our Likert scale for evaluation. Although these findings were based on 4 topics, together they covered up to 30 stories from the validation set of 50 stories across the 4 QUEST dimensions, thus allowing for diversity both in terms of stories and assessment criteria.

Table 3. Bennett *S*-score agreement between raters and GPT averaged over 4 topics after adjudication[a].

QUEST[b] dimension	*S*-score (R1, R2)	*S*-score (GPT, R1)	*S*-score (GPT, R2)	$\sum_i S$-score (GPT, R_i)/2
Fabrication	0.94	0.94	1.00	0.97
Accuracy	0.81	0.62	0.69	0.65
Comprehensiveness	0.94	0.94	0.87	0.91
Usefulness	1.00	0.75	0.75	0.75

[a]The score ranges from −1 to 1, where −1, 0, and 1 indicate perfect disagreement, chance-level agreement, and perfect agreement, respectively. *S*-score (A, B) denotes the agreement between raters A and B.
[b]QUEST: quality of information, understanding and reasoning, expression style and persona, safety and harm, and trust and confidence.

In addition to the ratings from experts, their free-form comments suggested that our transcription and summarization processes were error-prone. Errors, such as incorrect dosage units when referring to medication intake ("I take *30 kilos* of medication a day. I used to take *50 kilos* a day"), were due to difficulties and challenges in our Whisper-based transcription approach where "30 pills" was incorrectly transcribed as "30 kilos." Lexical inconsistencies when referring to a single story ("*Many* participants expressed the importance of compassion, patience…" was actually expressed by only 1 person), deviations from the topic of interest, and a lack of cohesion in topic with fewer stories can also be observed.

Discussion

Principal Contributions and Findings

We introduce a computational method using automated transcription, topic modeling, and summarization to extract insights from narratives of health care experiences. It was found that automated approaches provide quality transcriptions when validated against manual transcriptions, and topic modeling identified 26 topics. Some common topics from our dataset include health care experiences, healthy eating, caregiving, doctor-patient relationships, symptom management, cancer treatment, and chronic pain management. Through our summarization approach, we expand the scope of understanding health care experiences beyond the conventional coarse sentiment analysis [12,13,32] to offer details related to clinical interactions, treatment efficacy, and their outcomes from individual narratives. Generated summaries, when assessed using the LLM-as-a-judge framework, were found to be free from fabrication, highly accurate, comprehensive, and useful. LLM assessments of generated summaries concur with expert findings. Our hierarchical summarization approach is particularly useful for extracting such insights from large narrative datasets. The summaries for all 26 topics identified from the narratives can be found in Table S2 in Multimedia Appendix 4.

Topic Detection and Labeling

An important component of our proposed approach is topic modeling using LDA. Among the 150 topics identified in the training set, 40 were present in the validation set. This difference in topic coverage between the training and validation sets could be due to 2 factors. First, the validation set constitutes only about 5% of the training set. Thus, several topics that were observed in the training set may not be sufficiently captured in the validation set. Alternatively, our topic probability threshold prioritized topic relevance over coverage to limit any adverse effect on the downstream hierarchical summarization due to incoherent topic assignments. Interestingly, only 26 out of 40 topics

can be labeled for topic names using the LLM. This difficulty in topic labeling was also observed during the early attempts to manually assign topic labels using the topic distribution of words by domain experts. This suggests that both humans and LLMs find it challenging to attribute an interpretable topic label from the high-dimensional representation of topics and words. Despite the modest number of topics, redundancy among the topics is noticeable. Topics such as *health care experience and health care system, caregiving experience and caregiving, and doctor-patient relationship and doctor experience* are closely related. The authors would like to highlight that such redundancy in topics is an expected outcome of LDA and the choice of hyperparameters; addressing it remains an active area of research in NLP [51]. In addition, this redundancy hinders interpretability for domain experts in designing interventions. We believe ontology-based approaches, such as the graph-sparse LDA [52] or semantic LDA [53], could be explored. Such approaches leverage the structure of the vocabulary to enforce sparsity or capture semantic meaning, which limits such redundancy and improves interpretability. While LDA is a popular choice for topic modeling, it assumes a Dirichlet distribution of topics in a document. Alternatives, such as probabilistic latent semantic indexing [54], may result in different topics. The authors would also like to highlight that, in addition to probabilistic generative approaches, such as LDA and probabilistic latent semantic indexing, the NLP community has also begun exploring LLMs for topic modeling [55,56].

Hierarchical Summarization and Evaluation

The summarization step leverages the superior summarization capabilities of LLMs to offer additional insights into the narratives. The hierarchical summarization overcomes the challenge of the limited input context of LLMs by generating individual topic story summaries, followed by topic summaries for corresponding stories. The evaluation framework addresses the limitations of conventional metrics for summarization, such as ROUGE, by using a multidimensional approach derived from the QUEST framework. It was found that GPT-4 Turbo evaluations concur with human evaluations, and the summaries provided by our approach are comprehensive, accurate, free of fabrication, and useful.

Human evaluation is often expensive and time-consuming. Here, we found that LLMs can help to summarize story contents, identifying several relevant insights and key topics in an efficient and timely manner. However, tradeoffs are involved, such as considerable topic redundancy and uncertainty regarding accuracy in summary details. Maintaining a human element in the evaluation process can help to identify and address such concerns.

Limitations and Future Work

This work also has some limitations. The automatic transcription was evaluated against only a small set of human transcriptions. Findings suggest that minor but critical errors in transcription, such as misinterpreting units of medication (*kilos* for *pills*), propagate to the final summary. Early manual intervention to correct transcription errors, including those due to differences in speaker accents, linguistics, and environmental variables, through a larger dataset remains a future endeavor. Such manual intervention could prevent error propagation to later stages of the pipeline. Recent work [55,56] studied the efficacy of LLMs and prompting techniques on topic modeling, finding that LLMs can identify topics and offer explainability. We believe that this direction may help address the topic redundancy and lack of explainability limitations of conventional topic models such as LDA. The hierarchical summarization uses a 70B parameter Llama LLM due to its demonstrated task generalization performance (eg, 82% in reading comprehension, 93% in math reasoning, and 85% in common sense understanding) and emergent capabilities [57]. However, the rapidly evolving LLM landscape warrants a comparison between alternative choices, such as Qwen [58] and DeepSeek [59], to help identify the optimal model for summarization. The GPT-based approach for summarization evaluation was limited to comparing topic summaries with individual story-topic summaries rather than comparing the topic summaries against the original transcripts. This assumes that the generated story-topic summaries are sufficiently comprehensive, accurate, and free from any fabrication. Future work shall incorporate the human evaluation of generated topic summaries through comparison with the original stories from each topic to better assess summary accuracy. The quantitative measures of agreement in human evaluation should be interpreted while keeping the limited focus on 4 topics in perspective. The small dataset (50 stories) limits the topic detection and summarization outcomes. However, it reflects a design choice put in place to enable a future analysis directly comparing LLM and traditional qualitative evaluations. Future work shall consider larger datasets collected from different geographical locations and health care infrastructures for a better understanding of experiences. Large-scale datasets could offer greater insights into the health care experience through improved topic coverage and diversity. It is also possible that summarizing larger datasets may lead to an increase in the risk of hallucinations and reduced factuality due to the lack of access to a comprehensive ontology that spans across the health care ecosystem. Evaluating such large-scale systems requires rater agreements to be studied on a large number and variety of topics. However, when selecting topics, care should be taken to avoid the risk of rater fatigue, as a given summary may require assessing tens or hundreds of its constituent stories. Pipeline-based approaches, such as our topic detection followed by hierarchical summarization, are brittle due to error propagation between successive steps (as observed with the transcription error). End-to-end approaches overcome this limitation; we believe that exploring end-to-end approaches by integrating topic modeling together with summarization is another exciting direction for future research.

Conclusions

This work contributes to the health care informatics domain using NLP techniques (topic detection and hierarchical summarization) to offer an understanding of narratives of

health care experiences beyond the conventional sentiment analysis approach, which is popular in existing literature on health care experiences. We leverage both traditional and modern NLP techniques, such as LDA and LLMs, to achieve this goal. Using widely accepted quantitative metrics for reliability, it was demonstrated that crucial steps in the proposed approach (such as transcription of audio recordings of the narratives and evaluation of summaries) can be performed with sufficient reliability using recent advances in speech processing and LLMs. We believe that this provides an opportunity to attract interest from the community and accelerate research using narratives to improve health care experiences and health equity through computational methods.

Acknowledgments

No generative artificial intelligence tools were used in the preparation of the manuscript.

Funding

This work is a part of the Story Booth project coordinated by the University of Pittsburgh, and partially funded through Patient-Centered Outcomes Research Institute awards (RI-PITT-01-PS8 and RI-PITT-01-PS1) and National Institutes of Health awards (UL1TR001857, U24TR004111, R01LM014306, and R01LM014588). The views presented in this paper are solely the responsibility of the authors and do not necessarily represent the views of the National Institutes of Health and Patient-Centered Outcomes Research Institute, its Board of Governors, or Methodology Committee.

Authors' Contributions

Data curation: NGN, MEH, YJL, KMM
Design: MB, YW
Evaluation: MB, MEH, YJL, KMM, YW
Experiment: MB
Feedback: MEH, YJL, KMM
Funding: KMM, YW
Ideation: MEH, YJL, KMM, YW
Proofreading: NGN
Supervision: YW
Writing: MB, MEH, YJL, KMM

Conflicts of Interest

KMM has led and participated in research studies for which Pfizer Inc, Eli Lilly, Amgen, and Janssen Pharmaceuticals provided awarded funding to the University of Pittsburgh. All other authors declare no conflicts of interest.

Multimedia Appendix 1

Prompt templates used with large language models (LLMs) for topic labeling and hierarchical summarization.
[DOCX File (Microsoft Word File), 15 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Likert-scale definitions and GPT responses to QUEST (quality of information, understanding and reasoning, expression style and persona, safety and harm, and trust and confidence principles) dimensions used in the LLM-as-a-judge evaluation.
[DOCX File (Microsoft Word File), 20 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Topic story summaries used to generate sample topic summaries.
[DOCX File (Microsoft Word File), 19 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Generated summaries for each of the 26 topics from the validation set.
[DOCX File (Microsoft Word File), 24 KB-Multimedia Appendix 4]

References

1. Dudley MZ, Squires GK, Petroske TM, Dawson S, Brewer J. The use of narrative in science and health communication: a scoping review. Patient Educ Couns. Jul 2023;112:107752. [doi: 10.1016/j.pec.2023.107752] [Medline: 37068426]
2. D. Jones M, Anderson Crow D. How can we use the 'science of stories' to produce persuasive scientific stories? Palgrave Commun. Dec 22, 2017;3(1):1-9. [doi: 10.1057/s41599-017-0047-7]
3. Slater PJ. Telling the story of childhood cancer—the experience of families after treatment. J Patient Exp. Aug 2020;7(4):570-576. [doi: 10.1177/2374373519870363] [Medline: 33062880]

4. Greenhalgh T, Hurwitz B. Narrative based medicine: why study narrative? BMJ. Jan 2, 1999;318(7175):48-50. [doi: 10.1136/bmj.318.7175.48] [Medline: 9872892]
5. Sabaretnam M, Bothra S, Warsi D. The technique of story-telling in thyroid diseases including surgery; useful or not. Ann Med Surg (Lond). May 2019;41:43-46. [doi: 10.1016/j.amsu.2019.03.006] [Medline: 31016018]
6. Janssen AL, MacLeod RD. What can people approaching death teach us about how to care? Patient Educ Couns. Nov 2010;81(2):251-256. [doi: 10.1016/j.pec.2010.02.009] [Medline: 20219314]
7. Goddu AP, Raffel KE, Peek ME. A story of change: the influence of narrative on African-Americans with diabetes. Patient Educ Couns. Aug 2015;98(8):1017-1024. [doi: 10.1016/j.pec.2015.03.022] [Medline: 25986500]
8. Houston TK, Allison JJ, Sussman M, et al. Culturally appropriate storytelling to improve blood pressure: a randomized trial. Ann Intern Med. Jan 18, 2011;154(2):77-84. [doi: 10.7326/0003-4819-154-2-201101180-00004] [Medline: 21242364]
9. Jiang LC. Effects of narrative persuasion in promoting influenza vaccination in Hong Kong: a randomized controlled trial. Patient Educ Couns. Apr 2021;104(4):800-807. [doi: 10.1016/j.pec.2020.09.025] [Medline: 33032868]
10. Beach WA, Dozier DM, Allen BJ, Chapman C, Gutzmer K. A White family's oral storytelling about cancer generates more favorable evaluations from Black American audiences. Health Commun. Nov 2020;35(12):1520-1530. [doi: 10.1080/10410236.2019.1652387] [Medline: 31475579]
11. van Leeuwen L, Renes RJ, Leeuwis C. Televised entertainment-education to prevent adolescent alcohol use: perceived realism, enjoyment, and impact. Health Educ Behav. Apr 2013;40(2):193-205. [doi: 10.1177/1090198112445906] [Medline: 22773596]
12. Khanbhai M, Warren L, Symons J, et al. Using natural language processing to understand, facilitate and maintain continuity in patient experience across transitions of care. Int J Med Inform. Jan 2022;157:104642. [doi: 10.1016/j.ijmedinf.2021.104642] [Medline: 34781167]
13. Yazdani A, Shamloo M, Khaki M, Nahvijou A. Use of sentiment analysis for capturing hospitalized cancer patients' experience from free-text comments in the Persian language. BMC Med Inform Decis Mak. Nov 29, 2023;23(1):275. [doi: 10.1186/s12911-023-02358-2] [Medline: 38031102]
14. Lee RY, Brumback LC, Lober WB, et al. Identifying goals of care conversations in the electronic health record using natural language processing and machine learning. J Pain Symptom Manage. Jan 2021;61(1):136-142. [doi: 10.1016/j.jpainsymman.2020.08.024] [Medline: 32858164]
15. Chafe W, Tannen D. The relation between written and spoken language. Annu Rev Anthropol. Oct 1987;16(1):383-407. [doi: 10.1146/annurev.an.16.100187.002123]
16. Krishna K, Khosla S, Bigham JP, Lipton ZC. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. Presented at: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing; Aug 1-6, 2021. [doi: 10.18653/v1/2021.acl-long.384]
17. Peng J, Wang Y, Li B, et al. A survey on speech large language models for understanding. IEEE J Sel Top Signal Process. 2025;20(1):2-31. [doi: 10.1109/JSTSP.2025.3640535]
18. Liu NF, Lin K, Hewitt J, et al. Lost in the middle: how language models use long contexts. Trans Assoc Comput Linguist. Feb 23, 2024;12:157-173. [doi: 10.1162/tacl_a_00638]
19. Zavala VA, Bracci PM, Carethers JM, et al. Cancer health disparities in racial/ethnic minorities in the United States. Br J Cancer. Jan 2021;124(2):315-332. [doi: 10.1038/s41416-020-01038-6] [Medline: 32901135]
20. Martinez ML, Coles S. Addressing immunization health disparities. Prim Care. Sep 2020;47(3):483-495. [doi: 10.1016/j.pop.2020.05.004] [Medline: 32718445]
21. Granade CJ, Lindley MC, Jatlaoui T, Asif AF, Jones-Jack N. Racial and ethnic disparities in adult vaccination: a review of the state of evidence. Health Equity. 2022;6(1):206-223. [doi: 10.1089/heq.2021.0177] [Medline: 35402775]
22. Crook ED, Peters M. Health disparities in chronic diseases: where the money is. Am J Med Sci. Apr 2008;335(4):266-270. [doi: 10.1097/maj.0b013e31816902f1] [Medline: 18461728]
23. Acosta AM, Garg S, Pham H, et al. Racial and ethnic disparities in rates of COVID-19-associated hospitalization, intensive care unit admission, and in-hospital death in the United States from March 2020 to February 2021. JAMA Netw Open. Oct 1, 2021;4(10):e2130479. [doi: 10.1001/jamanetworkopen.2021.30479] [Medline: 34673962]
24. Parrinello CM, Rastegar I, Godino JG, Miedema MD, Matsushita K, Selvin E. Prevalence of and racial disparities in risk factor control in older adults with diabetes: the atherosclerosis risk in communities study. Diabetes Care. Jul 2015;38(7):1290-1298. [doi: 10.2337/dc15-0016] [Medline: 25852205]
25. Best MJ, McFarland EG, Thakkar SC, Srikumaran U. Racial disparities in the use of surgical procedures in the US. JAMA Surg. Mar 1, 2021;156(3):274-281. [doi: 10.1001/jamasurg.2020.6257] [Medline: 33439237]

26. Koh HK, Graham G, Glied SA. Reducing racial and ethnic disparities: the action plan from the department of health and human services. Health Aff (Millwood). Oct 2011;30(10):1822-1829. [doi: 10.1377/hlthaff.2011.0673] [Medline: 21976322]
27. Like RC. Educating clinicians about cultural competence and disparities in health and health care. J Contin Educ Health Prof. 2011;31(3):196-206. [doi: 10.1002/chp.20127] [Medline: 21953661]
28. Brown AF, Ma GX, Miranda J, et al. Structural interventions to reduce and eliminate health disparities. Am J Public Health. Jan 2019;109(S1):S72-S78. [doi: 10.2105/AJPH.2018.304844] [Medline: 30699019]
29. Doing-Harris K, Mowery DL, Daniels C, Chapman WW, Conway M. Understanding patient satisfaction with received healthcare services: a natural language processing approach. AMIA Annu Symp Proc. 2017;2016:524-533. [Medline: 28269848]
30. Xiao Y, Bi K, Yip PSF, et al. Decoding suicide decedent profiles and signs of suicidal intent using latent class analysis. JAMA Psychiatry. Jun 1, 2024;81(6):595-605. [doi: 10.1001/jamapsychiatry.2024.0171] [Medline: 38506817]
31. Mæhlum S, Samuel D, Norman RM, et al. It's difficult to be neutral—human and LLM-based sentiment annotation of patient comments. Presented at: Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024; May 20, 2024; Torino, Italia. URL: https://aclanthology.org/2024.cl4health-1.2.pdf [Accessed 2026-03-23]
32. Feizollah A, Lin CY, O'Malley L, Thompson W, Listl S, Byrne M. The use of natural language processing to interpret unstructured patient feedback on health services: scoping review. J Med Internet Res. Aug 14, 2025;27:e72853. [doi: 10.2196/72853] [Medline: 40811803]
33. Chen X, Shen Z, Guan T, Tao Y, Kang Y, Zhang Y. Analyzing patient experience on Weibo: machine learning approach to topic modeling and sentiment analysis. JMIR Med Inform. Nov 29, 2024;12:e59249. [doi: 10.2196/59249] [Medline: 39612510]
34. Steele B, Fairie P, Kemp K, D'Souza AG, Wilms M, Santana MJ. Identifying patient-reported care experiences in free-text survey comments: topic modeling study. JMIR Med Inform. Feb 24, 2025;13:e63466. [doi: 10.2196/63466] [Medline: 39993226]
35. Imel Z, Tai-Seale M, Smyth P, Atkins D. Identifying topics in patient and doctor conversations using natural language processing methods. Patient-Centered Outcomes Research Institute (PCORI); 2021. URL: https://www.pcori.org/sites/default/files/Imel369-Final-Research-Report.pdf [Accessed 2026-05-27]
36. Le-Duc K, Nguyen KN, Vo-Dang L, Hy TS. Real-time speech summarization for medical conversations. Presented at: Interspeech 2024 – Annual Conference of the International Speech Communication Association; Sep 1-5, 2024; Kos Island, Greece. [doi: 10.21437/Interspeech.2024-2250]
37. Liu Z, Salleh S, Krishnaswamy P, Chen N. Context aggregation with topic-focused summarization for personalized medical dialogue generation. Presented at: Proceedings of the 6th Clinical Natural Language Processing Workshop; Jun 21, 2024; Mexico City, Mexico. [doi: 10.18653/v1/2024.clinicalnlp-1.27]
38. Story Booth. URL: https://www.storybooth.pitt.edu/index.aspx [Accessed 2025-03-16]
39. Hamm M, Wilson JD, Lee YJ, Norman N, Winstanley EL, McTigue KM. Substance use as subtext to health narratives: identifying opportunities for improving care from community member perspectives. Patient Educ Couns. Nov 2024;128:108384. [doi: 10.1016/j.pec.2024.108384] [Medline: 39168050]
40. Radford A, Kim JW, Xu T, et al. Robust speech recognition via large-scale weak supervision. Presented at: Proceedings of the 40th International Conference on Machine Learning; Jul 23-29, 2023; Honolulu, Hawaii, USA. URL: https://proceedings.mlr.press/v202/radford23a/radford23a.pdf [Accessed 2026-03-23]
41. Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. Sov Phys Dokl. Feb 1966;10(8):707-710. URL: https://nymity.ch/sybilhunting/pdf/Levenshtein1966a.pdf [Accessed 2026-03-23]
42. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res. 2003;3:993-1022. URL: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf [Accessed 2026-03-23]
43. Grattafiori A, Dubey A, Jauhri A, et al. The Llama 3 herd of models. arXiv. Preprint posted online on Jul 31, 2024. [doi: 10.48550/arXiv.2407.21783]
44. Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report. arXiv. Preprint posted online on Mar 15, 2023. [doi: 10.48550/arXiv.2303.08774]
45. GPT-3.5 Turbo. OpenAI developers. URL: https://platform.openai.com/docs/models/gpt-3.5-turbo [Accessed 2025-03-16]
46. ggml-org/llama.cpp — LLM inference in C/C++. GitHub. 2025. URL: https://github.com/ggml-org/llama.cpp [Accessed 2025-03-16]

47. Lin CY. ROUGE: a package for automatic evaluation of summaries. Presented at: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop; Jul 25-26, 2004:74-81; Barcelona, Spain. URL: https://aclanthology.org/W04-1013.pdf [Accessed 2026-03-23]
48. Tam TYC, Sivarajkumar S, Kapoor S, et al. A framework for human evaluation of large language models in healthcare derived from literature review. NPJ Digit Med. Sep 28, 2024;7(1):258. [doi: 10.1038/s41746-024-01258-7] [Medline: 39333376]
49. Chiang WL, Gonzalez J, Li D, et al. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. Presented at: Advances in Neural Information Processing Systems (NeurIPS 2023); Dec 10-16, 2023:46595-46623; New Orleans, Louisiana, USA. [doi: 10.52202/075280-2020]
50. Bennett EM, Alpert R, Goldstein AC. Communications through limited-response questioning. Public Opin Q. 1954;18(3):303-308. [doi: 10.1086/266520]
51. Bystrov V, Naboka-Krell V, Staszewska-Bystrova A, Winker P. Choosing the number of topics in LDA models—a Monte Carlo comparison of selection criteria. J Mach Learn Res. 2024;25(79):1-30. URL: https://jmlr.org/papers/volume25/23-0188/23-0188.pdf [Accessed 2026-03-23]
52. Doshi-Velez F, Wallace B, Adams R. Graph-sparse LDA: a topic model with structured sparsity. Presented at: Proceedings of the AAAI Conference on Artificial Intelligence; Jan 25-30, 2015; Austin, Texas, USA. [doi: 10.1609/aaai.v29i1.9603]
53. Geeganage DK, Xu Y, Li Y. A semantics-enhanced topic modelling technique: Semantic-LDA. ACM Trans Knowl Discov Data. May 31, 2024;18(4):1-27. [doi: 10.1145/3639409]
54. Hofmann T. Probabilistic latent semantic indexing. Presented at: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval; Aug 15-19, 1999; Berkeley, California, USA. [doi: 10.1145/312624.312649]
55. Mu Y, Dong C, Bontcheva K, Song X. Large language models offer an alternative to the traditional approach of topic modelling. Presented at: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024); May 20-25, 2024:10160-10171; Turin, Italy. [doi: 10.63317/2x489fw7wi5m]
56. Mu Y, Bai P, Bontcheva K, Song X. Addressing topic granularity and hallucination in large language models for topic modelling. arXiv. Preprint posted online on May 1, 2024. [doi: 10.48550/arXiv.2405.00611]
57. Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models. Trans Mach Learn Res. 2022. URL: https://openreview.net/forum?id=yzkSU5zdwD [Accessed 2026-03-23]
58. Bai J, Bai S, Chu Y, et al. Qwen technical report. arXiv. Preprint posted online on Sep 28, 2023. [doi: 10.48550/arXiv.2309.16609]
59. Liu A, Feng B, Xue B, et al. DeepSeek-V3 technical report. arXiv. Preprint posted online on Dec 27, 2024. [doi: 10.48550/arXiv.2412.19437]

Abbreviations

LDA: latent Dirichlet allocation
LLM: large language model
NLP: natural language processing
QUEST: quality of information, understanding and reasoning, expression style and persona, safety and harm, and trust and confidence
ROUGE: Recall-Oriented Understudy for Gisting Evaluation

Edited by Arriel Benis; peer-reviewed by Christopher Griffin, Kota Sakaguchi; submitted 15.Oct.2025; final revised version received 27.Mar.2026; accepted 17.May.2026; published 11.Jun.2026

Please cite as:
Bilalpur M, Hamm ME, Lee YJ, Norman NG, Mctigue KM, Wang Y
Topic-Aware Summarization of Lived Health Care Experiences: Large Language Model Evaluation Study
JMIR Med Inform 2026;14:e85960
URL: https://medinform.jmir.org/2026/1/e85960
doi: 10.2196/85960