

Comparative Analysis of Narrative Generation Techniques for Lifelog Data Summarization

1st Kaushal Mishra
School of Computing
Dublin City University
Dublin, Ireland
kaushal.mishra2@mail.dcu.ie
Student ID: 23267061

2nd Logeshwaran Inbaraj
School of Computing
Dublin City University
Dublin, Ireland
logeshwaran.inbaraj2@mail.dcu.ie
Student ID: 23269893

Abstract—Lifelogging involves the continuous collection of personal data through sensors and digital tools, generating vast amounts of multi-modal data. Summarizing this data into coherent and engaging narratives is essential for personal reflection, health monitoring, and goal tracking. Traditional summarization methods like template-based and rule-based generation offer structured but inflexible narratives, while statistical models lack nuanced understanding. Recent advancements in Large Language Models (LLMs) provide a promising alternative due to their superior language comprehension and generation capabilities. This study compares four narrative generation techniques—template-based, rule-based, statistical machine learning, and LLM-based—using lifelog data. Evaluation criteria include accuracy, relevance, coherence, and engagement. Our findings demonstrate that LLM-based method outperforms others, offering highly accurate, relevant, and engaging summaries. This research underscores the potential of LLMs in lifelog data summarization and suggests future exploration of hybrid approaches for enhanced performance.

Index Terms—lifelog, lifelogging, summarization, LLM

I. INTRODUCTION

A. Background

Lifelogging refers to the continuous capture of various data points related to an individual's daily life through the use of sensors, wearables, and other digital tools. This data can include geographical locations, physiological measurements, activity records, and other personal experiences. With the advent of wearable devices and smartphones, lifelogging has become increasingly popular as these technologies enable the capture of various aspects of an individual's daily life [2].

These lifelog data can be leveraged to generate narratives that provide meaningful insights into one's activities [3] [4] [5].

Over the years, lifelogging technology has evolved significantly, providing richer data sources for analysis and narrative generation [20]. The applications of lifelogging extend to health monitoring, personal analytics, and memory augmentation. However, the vast amount of data generated by lifelogging devices poses significant challenges in terms of data management, analysis, and utilization. Summarizing lifelog data into meaningful narratives is crucial for making this data comprehensible and useful. Summaries can help individuals reflect on their daily activities, track their personal goals, and share their experiences in a structured manner. Nevertheless, generating coherent and relevant summaries from raw lifelog data is complex due to its continuous, multi-dimensional, and often unstructured nature.

logging devices poses significant challenges in terms of data management, analysis, and utilization. Summarizing lifelog data into meaningful narratives is crucial for making this data comprehensible and useful. Summaries can help individuals reflect on their daily activities, track their personal goals, and share their experiences in a structured manner. Nevertheless, generating coherent and relevant summaries from raw lifelog data is complex due to its continuous, multi-dimensional, and often unstructured nature.

B. Problem Statement

The primary challenge in summarizing lifelog data is creating narratives that are not only accurate and relevant but also engaging and personalized. Traditional methods like template-based story generation and rule-based natural language generation (NLG) have been used with varying degrees of success. However, these methods often lack flexibility and creativity. Statistical machine learning models offer another approach but might miss the nuanced understanding required for high-quality summaries. Recent advancements in Large Language Models (LLMs) present a promising alternative, potentially offering superior summarization capabilities due to their advanced language understanding and generation skills.

C. Objectives

The main objectives of this practicum are: Generate summaries using various methods for narrative generation: Template-Based Story Generation, Rule-Based Natural Language Generation (NLG), Statistical Machine Learning Models, and LLM-Based Summary Generation. Evaluate these methods to demonstrate that LLM-based summary generation is superior in terms of accuracy, relevance, coherence and engagement.

D. Scope of the Study

This study will explore the application of these four different narrative generation methods on lifelog data. Each method will be implemented and evaluated based on predefined criteria.

The effectiveness of each approach will be assessed through both quantitative metrics and qualitative user feedback. The ultimate goal is to identify the most effective method for summarizing lifelog data, with a particular focus on the potential advantages of LLM-based generation.

II. LITERATURE REVIEW

This section delves into the existing body of research relevant to lifelog data and narrative generation techniques. It explores the nature and challenges of lifelog data, examining various narrative generation methods, including template-based story generation, rule-based natural language generation (NLG), statistical machine learning models, and Large Language Models (LLMs). Additionally, it discusses the metrics used to evaluate the effectiveness of these narrative generation techniques and review related work in the field to contextualize our study within the broader landscape of lifelog data summarization.

A. Lifelog Data and Its Challenges

Lifelogging, as discussed by Gurrin, Smeaton, and Doherty [2], provides a comprehensive understanding of personal big data, which is essential for generating meaningful narratives. The advent of deep learning techniques has further revolutionized the field [17] [14].

Lifelogging generates vast amounts of continuous, multi-modal data encompassing various aspects of an individual's daily life. This data typically includes images, timestamps, geographical locations, and biometric sensor readings such as heart rate and activity levels. The diverse nature of lifelog data presents several significant challenges in terms of curation, storage, and analysis.

1) *Types of Content*: The primary content types in lifelog data are images and extensive metadata. Images capture moments of an individual's daily activities, while metadata includes detailed information such as location coordinates, timestamps, and biometric sensor data (e.g., heart rate, steps taken, and calories burned). This combination of visual and contextual data provides a comprehensive view of daily life but also complicates data management and analysis.

2) *Challenges in Curating and Storing Data*: Curating and storing lifelog data is complex due to its volume and variety. The continuous capture of high-resolution images results in significant storage requirements. Additionally, metadata, which includes detailed contextual information, needs to be stored in a structured manner for efficient retrieval and analysis. Ensuring data integrity and privacy is another critical challenge, as lifelogging often involves sensitive personal information. [2].

3) *Challenges in Accessing and Analyzing Data*: Accessing and analyzing lifelog data to create meaningful summaries or narratives involves several hurdles:

- **Volume and Complexity**: The sheer volume of data makes it difficult to process and analyze efficiently. The continuous, multi-dimensional nature of lifelog data adds to its complexity.

- **Data Integration**: Integrating image data with metadata requires sophisticated techniques to derive meaningful insights. For instance, making sense of images often involves using computer vision techniques to identify objects and activities within the images.
- **Generating Meaningful Summaries**: Creating coherent and relevant narratives from raw lifelog data is challenging due to the unstructured nature of the data. Effective summarization requires advanced methods to interpret the data accurately and generate engaging narratives.

B. Narrative Generation Techniques

Several studies have examined different approaches to narrative generation from lifelog data, each with its advantages and limitations [3] [4] [5] [15] [16] [17]. For instance, research has highlighted the role of personal lifelogs in fostering self-reflection and memory retrieval. Gurrin, Smeaton, and Doherty [2] discuss lifelogging as personal big data, which is crucial for understanding the data foundation required for narrative generation and emphasize the significance of meaningful summarization in enhancing personal memory and reflection. Another key area of exploration is the application of Natural Language Generation (NLG) technologies to lifelog data. Doherty and Smeaton [3] investigate how NLG can transform raw lifelog data into coherent and readable summaries.

The further subsections discuss about the different methods which have been used in the project for summarising lifelog metadata and generating summaries. The evaluation of these techniques has also been discussed.

1) *Template-Based Story Generation*: Template-based story generation involves using predefined templates to create narratives. These templates are filled with extracted lifelog data, such as activities, locations, and times. This method ensures consistency and coverage of key data points but often lacks flexibility and creativity [22]. Template-based approaches are efficient in generating structured narratives but may produce repetitive and less engaging stories [3].

2) *Rule-Based Natural Language Generation (NLG)*: Rule-based NLG creates narratives based on a set of predefined rules and logic. This method allows for controlled storytelling and the implementation of complex narrative logic based on activity types, sequences of events, and other data-driven conditions. While rule-based NLG can produce detailed and accurate summaries, it requires extensive setup and maintenance, making it less robust and adaptive compared to other methods [4].

3) *Statistical Machine Learning Models*: Statistical machine learning models generate summaries by identifying patterns and relationships within the data. Techniques such as clustering and regression models can be used to predict events and summarize lifelog data based on statistical significance. While this approach offers a different analytical perspective, it may lack the nuanced understanding and narrative generation capabilities of more advanced models like LLMs [5].

4) *Large Language Models (LLMs)*: LLMs, such as GPT-3 and its successors, leverage deep learning techniques to

generate human-like text based on vast amounts of training data. LLMs have demonstrated remarkable capabilities in various natural language processing (NLP) tasks, including text summarization, translation, and generation. These models can produce highly coherent, relevant, and engaging narratives from lifelog data, making them a promising tool for this application [6] [1]. The ability of LLMs to understand context and generate personalized content sets them apart from traditional methods.

C. Evaluation Metrics for Text Generation

The evaluation of text generation, especially narratives from lifelog data, is a critical aspect of research in this field. Effective evaluation ensures that generated summaries and narratives are useful, accurate, and engaging. Evaluating lifelog summaries is particularly important because these summaries help transform raw, voluminous lifelog data into coherent and meaningful narratives that users can easily understand and reflect upon.

Summarizing lifelog data is inherently challenging due to the need to balance multiple factors such as the purpose of the summary, the coverage of content, and the fluency of the generated text. A good summary must accurately reflect the underlying data, cover the most important events, and be coherent and engaging. These complexities make the task of summarization particularly difficult and highlight the need for robust evaluation metrics.

Various papers address the metrics and methods for assessing the quality of generated narratives. Gatt and Krahmer [9] provide a comprehensive discussion on these evaluation metrics. Additionally, Schuster and Nakajima [15], Ren [16], and LeCun [17] have proposed various methods for narrative generation from lifelog data, each with distinct advantages and limitations. Notably, Gurrin, Smeaton, and Doherty [2] consider lifelogging as a form of personal big data, underscoring the foundational data requirements for effective narrative generation. The advancement of deep learning techniques has further transformed the field, as evidenced by the work of LeCun [17] and Young [14].

Evaluating the quality of generated text, especially narratives from lifelog data, involves multiple criteria:

- Accuracy: How accurately does the generated summary reflect the actual lifelog data?
- Relevance: Are the narratives relevant to the user's experiences and the underlying data?
- Coherence: How logically and fluently are the stories constructed?
- Engagement: How engaging and interesting are the narratives to the users?
- Robustness: How well can the method be applied to larger or different datasets?

Commonly used evaluation metrics include BLEU, ROUGE, and METEOR for automated assessment, while human evaluations provide insights into subjective aspects like coherence and engagement [7] [8] [9]. Gatt and Krahmer [9] provide a comprehensive survey on the state-of-the-art natural

language generation, highlighting the importance of these evaluation metrics. Additionally, embedding-based evaluation methods have shown promise in capturing semantic similarity more effectively [18] [19].

By employing these metrics, researchers can systematically assess the effectiveness of narrative generation methods, ensuring that the generated texts are not only accurate and relevant but also coherent and engaging for users. This evaluation is essential for advancing the field and developing tools that can transform lifelog data into valuable narratives.

III. METHODOLOGY

This section outlines the methodological framework employed in this study to generate narratives from lifelog metadata. We describe the dataset used, preprocessing steps, narrative generation techniques, and the evaluation metrics applied. We used Python and many of its libraries in Jupyter Notebook and Google Colab for this research. The methodology section details the processes and algorithms implemented to transform raw metadata into coherent narratives and assesses the performance of these methods. Fig 1 given below shows a concept map that will help with explainability. Fig 1 shows a concept map which helps in understanding this project.

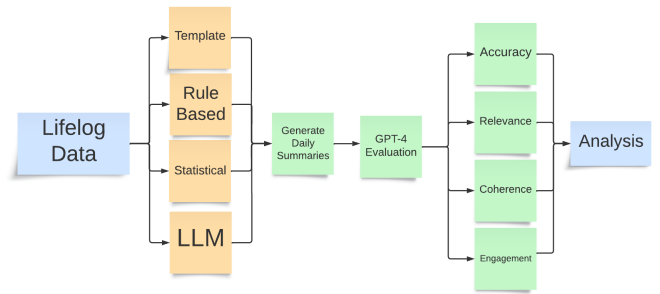


Fig. 1. Concept map for this research

A. Dataset

1) *Description*: Three datasets were provided by Cathal Gurrin from the ADAPT Centre, DCU, Ireland: the Lifelog image dataset, the Visual Concepts dataset using visual embeddings, and the lifelog metadata dataset. This research focuses exclusively on the metadata file, despite having access to both the image and visual concepts datasets. The aim is to explore the potential of generating narratives based solely on structured contextual information, without relying on image analysis.

```

Dataset Overview:
Total images in the dataset: 183,412
Average images per day: 1,609
Total minutes in a day: 1,440
Timelines covered:
2015-02-23 to 2015-03-20 (26 days)
2016-08-08 to 2016-10-05 (59 days)
2018-05-03 to 2018-05-31 (29 days)
Total days: 114
  
```

Listing 1. Dataset info

For this research, the 2015 and 2016 data were used as the training set, comprising 85 days, and the 2018 data were used as the testing set, comprising 29 days. It was observed that the 2018 data, or the testing set, had slightly higher calories and steps compared to the 2015 and 2016 training data. However, heart rate was consistent across both sets. This split was chosen to train on past data and test on future data rather than just splitting it randomly.

The lifelog metadata utilized in this project includes various types of personal data captured over an extended period. For more information about the metadata file used in this project, refer to the LSC data description provided by the Lifelog Search Challenge [21]. This dataset, along with more details, can be found on their official website ¹.

2) *Metadata File Details*: The lifelog data used in this project is derived from a comprehensive metadata file, which includes 164,160 entries, each representing a minute-by-minute record of lifelog data. The file, approximately 16.3 MB in size, includes the following columns:

- *minute_id*: A unique identifier for each minute.
- *utc_itimeandlocal_itime*: Timestamps indicating the time in UTC and local time zones.
- *timezone*: The time zone information.
- *lat* and *lon*: Latitude and longitude coordinates, providing geographical location data.
- *semantic_name*: Descriptive location names, such as "Home" or "Office".
- *elevation*: Elevation data, indicating the altitude at the given location.
- *speed*: The speed of movement at the given time.
- *heart*: Heart rate data, indicating the user's heart rate.
- *calories*: The number of calories burned.
- *activity_ttype*: The type of physical activity being performed, if available.
- *steps*: The number of steps taken.

This rich metadata provides a detailed view of an individual's daily activities, enabling the generation of comprehensive and meaningful narratives. By focusing on the metadata, key aspects of the user's day can be summarized without relying on image analysis, simplifying the data processing and narrative generation tasks.

3) *Data Preprocessing*: The following steps outline the data preparation process:

- **Data Import and Cleaning**: Import the lifelog data from CSV files, Clean the data by removing any missing or irrelevant entries, and Convert date and time fields to appropriate datetime formats
- **Feature Extraction**: Extract key features such as day of the week, year, and month from the datetime fields, Encode categorical variables like 'semantic name', 'activity type', 'day of week', and 'month' using Label Encoding.
- **Normalization**: Normalize numerical features such as heart rate, calories burned, steps taken, and activity time spent using StandardScaler.

- **Aggregation**: Aggregate data on a daily basis to get average heart rate, total calories burned, total steps, and total activity time spent per day. Concatenate unique semantic names and activity types with their respective total time spent.

The processed dataset was divided by days to facilitate the generation of daily summaries. Generating hourly summaries would have been time-consuming and the metrics captured in the dataset were not suitable for such detailed granularity.

These steps ensure that the data is in a consistent format, making it ready for use in narrative generation and subsequent evaluation.

B. Narrative Generation Methods and Implementation

In this subsection, four methods used for narrative generation are explained. They are Template based, Rule based, Statistical Machine Learning Based, and Large Language Model (LLM) based methods. Each method is discussed in detail to illustrate its contribution to the overall narrative generation process.

1) *Template-Based Method*: The template-based approach involved using predefined templates to generate summaries for the lifelog data. This method ensures consistency in the structure of the summaries and provides a straightforward way to present the daily activities recorded in the lifelog data. Steps:

- **Template Definition**: Many predefined templates were created for each column. The templates included placeholders for various data points such as date, locations visited, steps taken, heart rate, calories burned, and activities performed. Some examples:

```
time_templates = [
    "On {day_of_week}, {month} {day}, you were at places {semantic_names}.",
    "During the day of {day_of_week}, {month} {day}, you visited {semantic_names}.",
    "Throughout {day_of_week}, {month} {day}, you spent time at {semantic_names}.",
    "On {day_of_week}, {month} {day}, your locations included {semantic_names}.",
    "During {day_of_week}, {month} {day}, you were located at {semantic_names}." ...
]

steps_templates = [
    "You were quite active with {steps} steps.",
    "You managed to take {steps} steps.",
    "You accumulated {steps} steps.",
    "You took {steps} steps.",
    "You recorded {steps} steps.",
    "You logged {steps} steps.",
    "You made {steps} steps." ...
]

heart_rate_templates = [
    "Your heart rate was steady at {heart:.2f} bpm.",
    "Your heart rate maintained at {heart:.2f} bpm.",
    "Your heart rate held steady at {heart:.2f} bpm.",
    "Your heart rate was measured at {heart:.2f} bpm.",
    "Your heart rate registered at {heart:.2f} bpm." ...
]
```

Listing 2. Templates examples

- **Summary Generation**: The templates were filled with the actual data for each day using Python's string formatting. This ensured that each summary followed the same structure and included all relevant information.

¹http://lifelogsearch.org/lsc/2020/lsc_data/

- Output: The generated summaries provided a clear and structured overview of the daily activities, making it easy to understand the key metrics and events of each day.

This template-based approach provides a structured and consistent way to generate summaries of the lifelog data, ensuring that all relevant information is included and presented clearly.

2) *Rule-Based Method*: In rule-based approach, the summaries were created based on a set of predefined rules and logic. This method allows for controlled storytelling with the ability to implement complex narrative logic. It uses the same templates but adds rules to make the output better.

Steps:

- Rule Definition: Specific rules were defined to generate summaries, focusing on key metrics and significant activities. Examples of rules:

```
If the number of steps taken in a day exceeds a certain
threshold, highlight it as a "high activity" day.
If the average heart rate is above a certain level,
mention it as a "high heart rate" day.
If certain locations are visited, mention them
specifically in the summary.
```

Listing 3. Rules examples

- Summary Generation: The rules were applied to the aggregated data to generate summaries. Each summary was crafted based on the conditions met by the data for each day.
- Output: The generated summaries provided a detailed overview of the day's activities, emphasizing significant metrics and events based on the defined rules.

This rule-based approach allows for more dynamic and detailed summaries by highlighting significant events and metrics based on predefined conditions.

3) *Statistical Machine Learning Method*: In this statistical model, clustering was used to identify patterns and it generated summaries based on statistical significance and relationships within the data. The clustering-based approach involved finding out the best features using Principal Component Analysis (PCA), and grouping similar days together using the k-means clustering algorithm. This method helps in identifying patterns and generating summaries based on the characteristics of each cluster. It uses the same templates, but adds more information from clusters to make the output better.

Steps:

- Data Encoding: Categorical data, such as semantic names and activity details, were encoded to numerical values to facilitate clustering.
- Clustering: The k-means algorithm was used to cluster the days based on the encoded features. The numbers of clusters and features were finalised using the training dataset and implemented on test dataset. Four clusters were made using this approach.
- Cluster Analysis and Summary Generation: Each cluster was analyzed to identify common patterns. Summaries were generated based on the characteristics of each cluster.

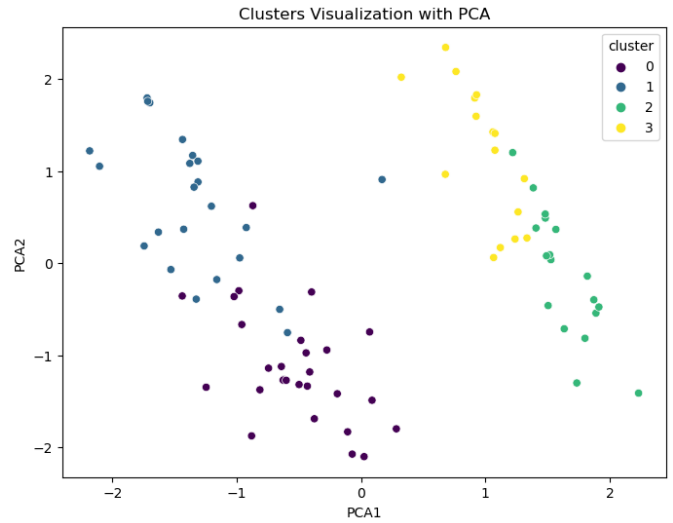


Fig. 2. Cluster Visualization using PCA

- Output: The generated summaries provided insights into the typical characteristics of each cluster, highlighting patterns and commonalities among similar days.

This clustering-based approach helps in identifying patterns and generating summaries based on the characteristics of each cluster, providing a deeper understanding of the lifelog data.

4) *LLM-Based Method*: The LLM-based approach utilizes OpenAI's GPT-4 model to generate summaries for the lifelog data. This method leverages the power of natural language processing to create coherent and engaging summaries based on detailed prompts. LLMs, such as GPT-4, are capable of producing coherent and contextually relevant narratives based on input data [1] [6] [10] [12]. Recent advancements in transformer architectures have significantly enhanced the capabilities of LLMs in generating high-quality text [11] [13].

Steps:

- OpenAI API Integration: API Key Setup: The OpenAI API key was loaded to authenticate and use the GPT-4 model.
- Prompt Creation: Detailed prompts were created to provide the model with sufficient context about the lifelog data for each day and we carefully selected the best prompt for testing by evaluating it for the training dataset. We first gave the API a context prompt then the summary prompt which generates the summary for each day.

```

$context_prompt$ = (
    "You are a helpful assistant. You will be
    provided with
    lifelog data that includes details like date, "
    "locations visited, steps taken, heart rate,
    calories burned,
    and activities. Your task is to generate "
    "a concise and coherent summary for each entry
    based on the
    provided data. Make sure the summaries are "
    "accurate, relevant, coherent, engaging, and have
    better
    transition between sentences for readability. "
    "Additionally, include a short personalized
    insight based on
    the data to enhance user engagement."
)

$prompt = (
    f"On {pd.to_datetime(row['day']).strftime('%A, %B
    %d')}, "
    f"You visited {row['semantic_name']}. "
    f"Steps taken: {int(row['steps'])}. "
    f"Heart rate: {row['heart']:.2f} bpm. "
    f"Calories burned: {row['calories']:.2f}. "
    f"Activities included: {row['activity_details']}. "
    "
    f"Summarize this data and add personalized
    insight with a motivational tone based on
    the data"
)

```

Listing 4. Context and Summary Prompts

- **Summary Generation:** The prompts were sent to the GPT-4 model to generate summaries.
- **Output:** The generated summaries provided a detailed and engaging overview of the daily activities, enriched with personalized insights and recommendations.

This LLM-based approach leverages the capabilities of advanced language models to generate detailed, coherent, and engaging summaries, providing valuable insights based on the lifelog data. It adds some insights after generating the summary for a day.

C. Outputs

Outputs of all methods for the same day:

```

Template: During Thursday, May 03, you visited places like
Home. You reached 198 steps. Your heart rate was steady
at 78.00 bpm. You spent your time on activities like
resting/idle (917 minutes), sleeping (465 minutes),
walking (2 minutes), transport (56 minutes).

Rule Based: Throughout the day on Thursday, May 03, you
visited Home. You walked a reasonable distance, taking
198 steps. Your heart rate was steady at 77.51 bpm.
Your day was highly active, burning 3019.77 calories.
Your activities for the day include: resting/idle (917
minutes), sleeping (465 minutes), walking (2 minutes),
transport (56 minutes).

Statistical(Clustering): On Thursday, May 03, you were seen
at Home. You managed to take 198 steps, maintaining a
moderate activity level. Your heart rate was 77.51 bpm.
Your high activity level resulted in 3019.77 calories
burned. Your primary activities were: resting/idle (917
minutes), sleeping (465 minutes), walking (2 minutes),
transport (56 minutes). This indicates a High Activity
, Moderate Heart Rate day.

LLM: "On Thursday, May 03, you spent the day at home. You
took 198 steps with a steady heart rate of 77.51 bpm
and burned 3019.77 calories. The majority of your day
was spent resting/idle (917 minutes), followed by

```

```

sleeping (465 minutes), with short periods of walking
(2 minutes) and transport (56 minutes).

```

Your body might have needed the rest and relaxation, as indicated by the high minutes spent resting and sleeping. However, incorporating more movement and activities into your day could contribute to a more balanced routine. Remember, every step counts towards a healthier lifestyle, so keep up the good work and find opportunities to stay active even while at home."

Listing 5. Outputs of all four methods

D. Evaluation

The evaluation phase is critical for assessing the performance and quality of the generated summaries. All summarization methods were evaluated on the testing set, and the results for the training and the testing sets were also compared.

1) *Metrics:* The evaluation focused on four key criteria: Accuracy, Relevance, Coherence, and Engagement.

Evaluation Criteria:

- **Accuracy:** This criterion measures how well the summary reflects the actual lifelog data. It assesses whether the key metrics and events are correctly represented in the summary. Scoring: Each summary was scored on a scale of 0 to 50, where 50 indicates perfect accuracy.
- **Relevance:** This criterion evaluates the pertinence of the summary to the provided lifelog data. It checks if the summary includes all significant activities and metrics relevant to the user's day. Scoring: Each summary was scored on a scale of 0 to 50, where 50 indicates perfect relevance.
- **Coherence:** This criterion measures the logical flow and readability of the summary. It assesses the structure, grammar, and clarity of the summary. Scoring: Each summary was scored on a scale of 0 to 50, where 50 indicates perfect coherence.
- **Engagement:** This criterion evaluates how engaging and interesting the summary is. It considers the narrative style, the use of engaging language, and the overall appeal of the summary. Scoring: Each summary was scored on a scale of 0 to 50, where 50 indicates perfect engagement.

2) *Methods:* Automated Evaluation using gpt-4o-mini: To ensure a consistent and unbiased evaluation, OpenAI's gpt-4o-mini model was used to assess the summaries based on the defined criteria. The following steps outline the process:

a. **Data Preparation:** The final summaries generated by each method were compiled into a single dataset. Each row in the dataset included the lifelog data, the generated summary, and metadata such as the date, day of the week, and other relevant details.

b. **Prompt Creation:** Detailed evaluation prompts were created to provide the model with context about the lifelog data and the generated summary.

Output prompt:

```

$prompt = (f"Here is the lifelog data: {lifelog_data}\n\n"
f"Here is the summary generated by the {
method_name} method: {summary}\n\n"
"Please evaluate this summary based on the
following criteria:\n"
"1. Accuracy: How accurately does this summary
reflect the lifelog data?\n"
"2. Relevance: Is the summary relevant to the
lifelog data?\n"
"3. Coherence: How logically coherent and
readable is the summary?\n"
"4. Engagement: How engaging and interesting is
the summary?\n"
"Please provide your response in the following
format:\n"
"Accuracy: <score>/50\n"
"Relevance: <score>/50\n"
"Coherence: <score>/50\n"
"Engagement: <score>/50\n"
"Explanation: <short explanation>")

```

Listing 6. Output Format Prompt

c. **Summary Evaluation:** The prompts were sent to the gpt-4o-mini model, which generated scores and explanations for each summary based on the defined criteria. A retry mechanism was implemented to handle potential rate limit errors from the API.

d. **Combining Results:** The evaluation results were combined into the final dataset, including the scores for Accuracy, Relevance, Coherence, and Engagement for each summary generated by the different methods.

Automated metrics like BLEU and ROUGE were not used in this research because we did not have a gold standard output that is used in these metrics and also the dataset used consisted mainly of numerical data, but these metrics are better suited for textual data, So they might not work well on our lifelog dataset.

Human evaluations were not involved in our project. we manually checked random LLM outputs just as a sanity check so that it worked as expected and didn't hallucinate.

IV. RESULTS AND ANALYSIS

This section presents the comparative analysis of the four narrative generation methods—Template-Based, Rule-Based, Clustering-Based, and LLM-Based—using the evaluation metrics of Accuracy, Relevance, Coherence, and Engagement. The performance of each method is assessed on both the training and testing datasets, as shown in Tables I and II.

TABLE I
AVERAGE EVALUATION SCORES BY METHOD TESTING SET

Method	Accuracy	Relevance	Coherence	Engagement
Template_based	44.9310	49.9310	46.4138	40.0000
Rule_based	49.8276	50.0000	45.4483	40.0000
Clustering_based	48.2759	50.0000	46.1379	40.4138
LLM_based	49.2414	50.0000	48.6552	44.2069

A. Template-Based Method

The Template-Based method, while straightforward, exhibited varying performance across different metrics.

Accuracy:

TABLE II
AVERAGE EVALUATION SCORES BY METHOD TRAINING SET

Method	Accuracy	Relevance	Coherence	Engagement
Template_based	45.3294	49.7176	45.9058	40.0470
Rule_based	49.6117	50.0000	45.4235	40.0235
Clustering_based	46.6705	49.8000	47.0000	40.3529
LLM_based	49.3176	50.0000	47.5294	42.3176

The Template-Based method achieved an accuracy score of 44.93 on the testing set and 45.33 on the training set (Tables I and II). This slight difference indicates that the method is relatively stable but limited by its rigid structure. Its accuracy is the lowest among the methods evaluated, reflecting the constraints of using predefined templates that may not capture the variability and nuances of the data.

Relevance:

In terms of relevance, the Template-Based method performed well, with scores of 49.93 on the testing set and 49.72 on the training set. These scores suggest that the templates are effective in ensuring the generated content is pertinent to the input data, though they may miss some subtleties captured by more flexible methods.

Coherence:

The coherence scores for the Template-Based method were 46.41 on the testing set and 45.91 on the training set. These scores are moderate, indicating that while the templates provide a logical structure, they may not produce as fluent narratives as more advanced methods.

Engagement:

The engagement scores were consistent, with 40 on the testing set and 40.05 on the training set. The method's rigid structure results in less engaging content, as it lacks the dynamic and varied nature of more sophisticated approaches.

B. Rule-Based Method

The Rule-Based method, leveraging predefined rules, showed strong performance in several areas.

Accuracy:

The Rule-Based method achieved the highest accuracy score among all methods, with 49.83 on the testing set and 49.61 on the training set. This high accuracy can be attributed to the precise rules that closely align with the input data, ensuring accurate content generation.

Relevance:

The relevance scores for the Rule-Based method were perfect, achieving 50 on both the testing and training sets. This indicates that the rules are highly effective in maintaining the relevance of the generated content.

Coherence:

However, coherence scores were lower, with 45.45 on the testing set and 45.42 on the training set. The deterministic nature of rule-based systems can lead to less fluid narratives, affecting overall coherence.

Engagement:

Engagement scores were 40 on the testing set and 40.02 on the training set. The method's rigid rules lead to more

monotonous and less engaging content compared to more flexible approaches.

C. Clustering-Based Method

The Clustering-Based method, which groups similar data points, exhibited mixed results.

Accuracy

The accuracy scores for the Clustering-Based method were 48.28 on the testing set and 46.67 on the training set. The slight decrease in accuracy on the testing set suggests variability within clusters, affecting performance.

Relevance

Relevance scores were high, achieving 50 on both the testing and training sets. This indicates that clustering effectively groups related data, ensuring relevant content generation.

Coherence

Coherence scores were 46.14 on the testing set and 47 on the training set. The method performed well, but the slight drop on the testing set indicates potential overfitting.

Engagement

Engagement scores were 40.41 on the testing set and 40.35 on the training set. The variability within clusters can impact engagement, though the method generally performs well.

D. LLM-Based Method

The LLM-Based method, using advanced language models, demonstrated exceptional performance.

Accuracy

The LLM-Based method achieved an accuracy score of 49.24 on the testing set and 49.32 on the training set. This consistency indicates the method's robustness and its ability to generalize well to unseen data.

Relevance

Relevance scores were perfect, with 50 on both the testing and training sets. The advanced language models ensure that the generated content is highly relevant to the input data.

Coherence

The LLM-Based method led in coherence, with scores of 48.66 on the testing set and 47.53 on the training set. The method's advanced language capabilities produce logically consistent and fluent narratives.

Engagement

Engagement scores were the highest among all methods, with 44.21 on the testing set and 42.32 on the training set. The dynamic and varied nature of the generated content makes it highly engaging.

Summary of Results

The detailed evaluation of each narrative generation method reveals distinct strengths and weaknesses:

Template-Based: Exhibits high relevance and moderate accuracy but falls short in coherence and engagement due to its rigid structure. Its performance is stable but limited by its predefined templates.

Rule-Based: Excels in accuracy and relevance, but lacks in coherence and engagement, limited by its deterministic rule

sets. It shows consistent performance and high relevance, but the narratives can be monotonous.

Clustering-Based: Shows high accuracy and relevance, but its coherence and engagement are somewhat impacted by cluster variability. The method performs well but shows slight overfitting issues.

LLM-Based: Demonstrates outstanding performance across all criteria, particularly in coherence and engagement, making it the most robust method among those evaluated. It generalizes well to new data and produces highly engaging content.

When comparing the results between the training and testing sets, we observe some minor differences across all methods. The Rule-Based and LLM-Based methods exhibit consistent performance with minimal variations, indicating their robust generalization capabilities to unseen data. The Clustering-Based and Template-Based methods, however, show slight declines in accuracy, coherence, and engagement on the testing set. These differences suggest potential overfitting, where the models perform well on the training data but struggle to maintain the same level of performance on new data. The rigid structure of the Template-Based method limits its adaptability, while the variability inherent in the clustering approach can lead to inconsistencies when applied to different datasets. These insights highlight the importance of evaluating methods on diverse datasets to ensure their robustness and generalization.

In summary, the comparison of test results with training results highlights the varying generalization abilities of the methods. The Rule-Based and LLM-Based methods consistently perform well, indicating strong generalization to new data. In contrast, the slight performance drops observed in the Clustering-Based and Template-Based methods on the testing set suggest potential overfitting. This reinforces the LLM-Based method as the most effective approach for lifelog data summarization, offering a balanced combination of accuracy, relevance, coherence, and engagement.

E. Notable Outputs

1) *Highest LLM score:* "On Wednesday, May 30, you had a balanced day with a visit to Costa Coffee and spending time at home. You took 6730 steps and burned 2859.47 calories. Your heart rate averaged at 86.78 bpm throughout the day. The activities you engaged in included resting/idle for 858 minutes, sleeping for 459 minutes, walking for 34 minutes, and using transport for 89 minutes.

Reflecting on your data, it's great to see the variety in your activities, from relaxation to a short walk and using transport. Remember, every step counts towards your well-being. Keep up the good work, and continue to incorporate movement into your daily routine for a healthy lifestyle."

this LLM response got these scores:

Accuracy: 50/50

Relevance: 50/50

Coherence: 50/50

Engagement: 45/50

2) *Lowest LLM score*: "On Monday, May 14, you had an active day moving around various locations like Collins Park, DCU School of Computing, and the Embassy of The People's Republic of China Visa & Consular Office before heading back home. You took a total of 12,576 steps, indicating a good amount of physical activity. Your heart rate averaged 77.36 bpm, and you burned 3702.74 calories throughout the day. The majority of your day was spent resting/idle, with significant time dedicated to sleeping and walking. Remember, each step taken and every calorie burned contributes to your overall well-being and health. Keep up the great work and stay active for a healthier lifestyle!".

this LLM response got these scores:

Accuracy: 45/50

Relevance: 50/50

Coherence: 45/50

Engagement: 42/50

Responses with the highest scores excel in accuracy, relevance, and coherence. They capture all lifelog data points accurately and present them in a well-organized, logical sequence. Despite their high engagement, they could benefit from a more personal touch or motivational insights. In contrast, responses with the lowest scores have minor inaccuracies, such as slight discrepancies in data representation, and less coherent structure. While still relevant, these summaries lack the clarity and engagement found in higher-scoring responses. Enhancing personalization, vivid language, and data precision can significantly improve lower-scoring responses. Generally the responses with lesser amount of text also have lower scores as we have observed.

V. CONCLUSION

This study presents a comprehensive analysis of narrative generation techniques for lifelog data summarization, comparing template-based, rule-based, statistical machine learning, and Large Language Model (LLM)-based methods. Lifelogging, with its continuous capture of personal data through sensors and digital tools, generates extensive multi-modal datasets that pose significant challenges for data management and analysis. Summarizing this data into coherent and engaging narratives is crucial for personal reflection, health monitoring, and goal tracking.

Our research evaluates each narrative generation technique using a lifelog dataset comprising metadata on geographical locations, physiological measurements, and activity records. The methods were assessed based on their accuracy, relevance, coherence, and engagement. Template-based and rule-based approaches, while structured, often lacked flexibility and creativity. Statistical machine learning models provided a different analytical perspective but fell short in understanding nuanced context. In contrast, LLM-based methods demonstrated superior performance, producing highly accurate, relevant, and engaging narratives due to their advanced language comprehension and generation capabilities. [1] [6]

Through rigorous evaluation, our findings underscore the potential of LLMs in enhancing the summarization of lifelog

data. The LLM-based method consistently outperformed others across all evaluation metrics, highlighting its ability to generate coherent and personalized narratives. This research underscores the transformative impact of LLMs on lifelog data summarization, paving the way for more sophisticated and user-centric applications in the future.

A. future works

Future work in lifelog narrative generation can enhance the techniques developed in this study by addressing several key areas. Integrating larger and more diverse datasets will improve model robustness and generalizability. Incorporating images and visual embeddings can create more engaging and contextually accurate narratives through a multi-modal approach.

Adopting advanced and better evaluation metrics can provide more accurate results for summary quality. Real-time data processing and narrative generation can enable dynamic and interactive lifelog applications.

Personalized and adaptive narrative strategies, tailored to individual preferences, will enhance user engagement. Finally, exploring the scalability and applicability of these methods to various lifelog data types, such as social media and online activity, will broaden their impact. These advancements will lead to more sophisticated, accurate, and user-centric lifelog narratives.

B. Figures and Tables

Figure 1: It is a concept. It helps in understanding the project structure. figure 2: It is a Graph showing the clusters formed using PCA. It is used in the Statistical method for generating summary.

Listing 1: It shows the information about the datasets involved in this project.

Listing 2: It shows some examples of the templates which are used in the template approach.

Listing 3: It shows some examples of rules which are used in the rule based approach.

Listing 4: It shows the context and summary prompts used in LLM approach. Context prompt gives the LLM some context and summary prompt gives the command to make the summary.

Listing 5: It shows the outputs of all four methods for 03 May, 2018. It is helpful in comparing the approaches at a glance.

Listing 6: It shows the prompt which is finally used to evaluate all the outputs.

Table 1: It shows the average evaluation scores for the testing dataset.

Table 2: It shows the average evaluation scores for the training dataset.

REFERENCES

- [1] Z. Wang, A. Gupta, "A Comparative Study on Narrative Generation Techniques Using Large Language Models," arXiv preprint arXiv:2407.00908, 2023.
- [2] C. Gurrin, A. F. Smeaton, A. R. Doherty, "Lifelogging: Personal big data," *Foundations and Trends in Information Retrieval*, vol. 8, no. 1, pp. 1-125, 2014.
- [3] A. R. Doherty, A. F. Smeaton, "Automatically segmenting lifelog data into events," in 2012 23rd Irish Signals and Systems Conference (ISSC), pp. 1-8, IEEE, 2012.
- [4] D. Gkatzia, O. Lemon, V. Rieser, "Natural language generation enhances human decision-making with uncertain information," arXiv preprint arXiv:1606.03254, 2016.
- [5] T. H. Nguyen, H. Le, "Clustering-based approach for summarizing daily lifelog data," *Expert Systems with Applications*, vol. 142, pp. 113008, Elsevier, 2020.
- [6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., "Language models are few-shot learners," arXiv preprint arXiv:2005.14165, 2020.
- [7] K. Papineni, S. Roukos, T. Ward, W. J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311-318, 2002.
- [8] C. Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text summarization branches out: Proceedings of the ACL-04 workshop*, vol. 8, pp. 74-81, 2004.
- [9] A. Gatt, E. Krahmer, "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation," *Journal of Artificial Intelligence Research*, vol. 61, pp. 65-170, 2018.
- [10] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, "Improving language understanding by generative pre-training," OpenAI, 2018.
- [11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, vol. 30, 2017.
- [12] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2019.
- [13] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38-45, 2020.
- [14] T. Young, D. Hazarika, S. Poria, E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55-75, 2018.
- [15] M. Schuster, K. Nakajima, "Japanese and Korean voice search," in 2012 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 5149-5152, IEEE, 2012.
- [16] S. Ren, K. He, R. Girshick, J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, vol. 28, 2015.
- [17] Y. LeCun, Y. Bengio, G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, 2015.
- [18] T. Mikolov, K. Chen, G. Corrado, J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.
- [19] D. P. Kingma, M. Welling, "Auto-encoding variational Bayes," arXiv preprint arXiv:1312.6114, 2013.
- [20] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, vol. 25, 2012.
- [21] Lifelog Search Challenge, "LSC2020 Dataset", [Online]. Available: http://lifelogsearch.org/lsc/2020/lsc_data/. [Accessed: Jul. 30, 2024].
- [22] C. van der Lee, E. Krahmer, and S. Wubben, "Automated learning of templates for data-to-text generation: comparing rule-based, statistical and neural methods," in *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, pp. 35-46, 2018. [Online]. Available: <https://aclanthology.org/W18-6504.pdf>.