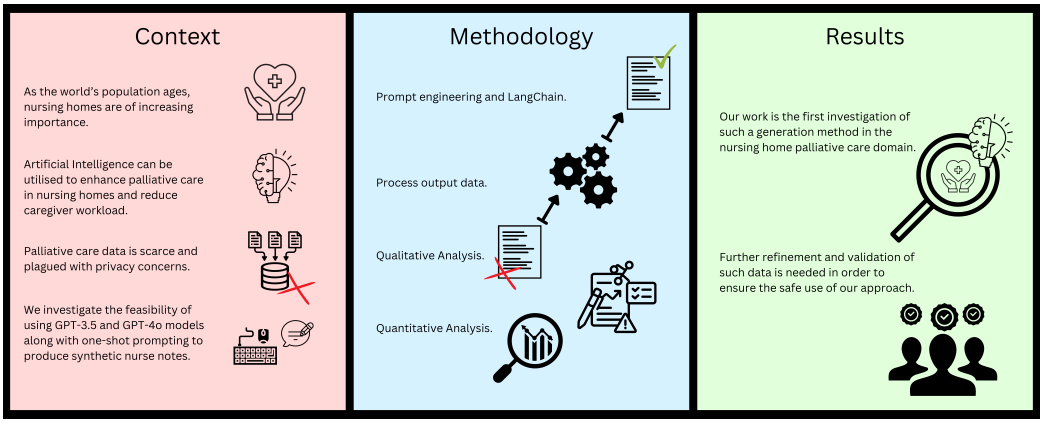
Graphical Abstract

A LangChain-Based Pipeline for One-Shot Synthetic Text Generation Using Generative Pre-Trained Transformers in Palliative Care Research

Isabel Ronan, Patrice Crowley, Eva Rombouts, Nicola Cornally, Mohamad M. Saab, David Murphy, Sabin Tabirca



A LangChain-Based Pipeline for One-Shot Synthetic Text Generation Using Generative Pre-Trained Transformers in Palliative Care Research

Isabel Ronan[a,1], Patrice Crowley[a], Eva Rombouts[b], Nicola Cornally[a], Mohamad M. Saab[a], David Murphy[a], Sabin Tabirca[a]

[a] *University College Cork, College Road, Cork, T12 K8AF, Cork, Ireland*
[b] *Rombouts Ouderengeneeskunde, Amsterdam, Netherlands*

Abstract

Objective: As the world's population ages, nursing homes are of increasing importance. In order to care for a growing number of older adults, intelligent technologies are needed. Artificial Intelligence can be utilised to enhance palliative care in nursing homes. However, the data needed to train artificially intelligent agents is lacking within this sensitive domain due to privacy issues. Therefore, it is difficult for researchers to develop technological solutions. With the advent of large language models, such as ChatGPT, new text generation methods are made possible using limited data. In this pilot study, we investigate the use of large language models to generate synthetic data.

Methods: We investigate the feasibility of using GPT-3.5 and GPT-4o models along with one-shot prompting to produce synthetic nurse notes which faithfully describe nursing home residents with met or unmet palliative care needs. We used LangChain to create a repeatable pipeline which can be adapted to different use-cases. We also compare the performance of both models using a set of qualitative and quantitative evaluations to determine

[1]Corresponding Author Email: i.ronan@cs.ucc.ie. Address: Western Gateway Building, University College Cork, Western Road, Cork, T12 XF62

which set of notes is more suitable for subsequent research.

Results: GPT-3.5 performed slightly better than GPT-4o in our qualitative healthcare professional analysis. Quantitative analysis revealed appropriately heterogenous results across contextual similarity, lexical overlap, sentiment, and readability scores.

Conclusion: Our work is the first investigation of such a generation method in the nursing home palliative care domain. Further refinement and validation of such data is needed in order to ensure the safe use of our approach.

Keywords: Synthetic, Palliative, Dataset, GenAI, LLM, Prompts, LangChain

1. Introduction

The world's population is ageing rapidly. According to UN predictions, over 60 countries will have over 2 million people aged 65 or older by 2030 [1]. A significant portion of older people reside in nursing homes due to chronic diseases or other serious ailments that require constant and specialised medical care [2]. Therefore, nursing home care is growing in importance as the global population ages [3].

It is becoming increasingly difficult for limited healthcare teams to provide high quality care in a timely fashion to a growing number of older adults using traditional methods. While such care has been shown to lead to increased quality of life, the ageing population is putting pressure on resource-constrained staff within nursing homes [4]. Older adults in these facilities tend to be reliant on a small number of staff members for their care [5]. New forms of care are needed in order to ensure nursing home residents are adequately cared for.

Palliative care is an approach to care which emphasises relief from suffering and improved quality of life [6, 7, 8]. Palliative approaches have been shown to improve the quality of care received by residents in nursing homes by reducing hospital admissions, increasing clinical care access when neces-

sary, and ameliorating family care perceptions [9]. Palliative care can be used to meet the diverse needs of older adults with appropriate impact and efficiency [10]. This type of care is becoming increasingly more popular as our society's health needs change.

Additionally, Artificial Intelligence (AI) can be utilised to improve the quality of life of nursing home residents, by providing health professionals with tools to alleviate workload. AI is a broad field encompassing all systems which exhibit human-like intelligence [11]. Artificially intelligent palliative care solutions involving machine learning (ML) have already been successful in other areas, such as in cancer patient treatment [12]. ML is a subset of AI, which is often used synonymously with the broader discipline [11]. ML uses algorithms coupled with large amounts of data to "learn" patterns and extract meaningful information from unseen input [11]. However, researchers must think carefully about the correct data to use for ML solutions in palliative care research.

Palliative care is highly subjective and personalised. Subtle changes in residents requiring increased care are difficult to identify. Nurse notes are subjective, personalised, human-based records of resident status which are collected across clinical settings and countries [13, 14]. Nurse notes may be utilised to create ML models for palliative care. While ML solutions using these notes could be hugely useful within nursing homes, data restrictions present issues for those training algorithms in the field [15]. Furthermore, a considerable quantity of nurse notes is required to train a model of sufficient reliability.

Due to uncertainty surrounding what data to share, insufficient infrastructure and non-standardised multi-modal methods of data collection, ML solutions are stagnating within the palliative nursing home domain [15]. Healthcare data is subject to intense privacy protection measures due to its highly sensitive nature [16]. Additionally, palliative care data faces further data anonymization issues due to its individualised, sparsely available nature [17].

Specifically, access to real notes from nursing homes is difficult to obtain as they are extremely personal and oftentimes unique to residents. Therefore, in order to test new approaches to machine learning within the space, a new method of obtaining big data, such as synthetic data generation (SDG), is required.

SDG is the creation of artificial data [16]. SDG circumvents many privacy and anonymization issues provided that no real patient data is used in the generation process. However, synthetic data must be plausible within the context it is generated for [16]. SDG in healthcare is fast becoming an attractive research area [18]. However, a thorough search of the literature revealed that no SDG procedures for palliative care exist, especially those for clinical note creation [17].

In this pilot study, we investigate the feasibility of using Generative Pre-Trained Transformer (GPT) models and one-shot prompting to generate synthetic notes which are representative of residents with met and unmet palliative needs in nursing homes. To generate the notes, we used LangChain, the OpenAI Application Programming Interface (API), prompt engineering techniques, and nurse-created example input. We also present the first performance comparison of GPT-3.5 and GPT-4o models for such data using both quantitative metrics and qualitative analysis by a subject matter expert. Our results show that such data needs further refinement and validation before use in the sensitive palliative care field. Additionally, comparison between synthetic and real notes is needed. The rest of this paper is divided into the following sections: literature review, methodology, results, discussion, and conclusions.

Statement of Significance

Problem or Issue	There is a lack of palliative care patient data from nursing home settings due to privacy concerns and the highly sensitive nature of end-of-life care. Lack of data limits artificial intelligence usage in these settings.
What is Already Known	GPT models can generate promising data for healthcare settings. Large language models pose problems when trained on sensitive data. Appropriate model inputs and repeatable processes are needed to generate usable synthetic data. There is a need to investigate free-text note generation specifically for palliative nursing home settings.
What this Paper Adds	This paper introduces a novel approach to palliative care nursing home data generation and evaluation using LangChain, GPT-3.5, and GPT-4o, along with both qualitative and quantitative evaluation measures.
Who would benefit from the new knowledge in this paper	Palliative care researchers, computer science researchers and healthcare professionals who have an interest in developing data-driven solutions to problems within long-term care settings.

2. Related Work

While clinical notes are crucial in many care settings, the highly individualised, subjective, free-text nature of them makes generation a unique challenge for researchers [19]. Clinical notes differ from relational data in the generation process due to the need for specialized natural language processing techniques [19]. However, as of yet, Electronic Health Records (EHR) text generation is still in its early stages [20]. Deep generative models, such as Generative Adversarial Networks (GAN), seem to be the most promising for palliative healthcare data generation at present; however, may of these

models tend to focus solely on tabular data creation [17].

Begoli et al. [21] proposed "SynthNotes", a generator framework for high-volume, high fidelity synthetic health notes. This framework operates on a template-filling principle using an openly available database for information extraction. Evaluation metrics for the same included BLEU and METEOR scores. This system was shown to be realistic and efficient; however, it only focused on the generation of mental health notes and was not extended for other use-cases, such as palliative care.

Libbi et al. [22] proposed the use of large language models (LLMs) to generate synthetic clinical notes. They used GPT-2 and real patient records to produce coherent notes, comparing these notes to alternative Long Short-Term Memory (LSTM) methods. A variety of evaluation metrics were used, such as ROUGE-N and human evaluation, to assess the quality of their synthetic data. Using in-text annotation, the models could produce artificial text that was automatically prepared for named entity recognition (NER) tasks. However, they highlight privacy concerns in their results; the model reproduced substantial portions of identifying information from the real records. Additionally they describe the need for further research before EHRs can be used as an anonymous alternative to real text.

Amin-Nejad et al. [23] used a vanilla transformer architecture to produce hierarchical long passage text when trained on a large dataset. However, they found the same model to be ineffective with smaller datasets. They compared these results against a GPT-2 model and found that GPT-2 performed better in resource-constrained environments with smaller amounts of data. They also found that GPT-2 was better suited to conditional language modelling. They used perplexity, BLEU and ROUGE scores to intrinsically evaluate the output text. However, they used a dataset which consisted solely of notes from critical care hospital admissions, thereby limiting the application of their approach to palliative care settings.

Sufi [24] has outlined an automated pipeline for generating synthetic hos-

pital data using GPT-4. This work is the first which aims to use the OpenAI API to create synthetic patient discharge messages. Results from these experiments indicate that GPT models can be used to produce synthetic data which is true to patient realities. However, the use of Microsoft Power Automate in this method limits scalability; only 70 patient messages were generated during this study. Additionally, this work does not focus on any of the subtleties associated with palliative care documentation.

Suvalov et al. [25] investigated the use of a locally-hosted GPT-2 model to generate synthetic data for further NER tasks. Their methods involved the use of approximately 10 million texts from 200,00 Estonian patients to generate 4100 documents for subsequent use in third-party models. However, their methodology does not guarantee that input data is not leaked from the synthetic data generation stages to the actual online models. Additionally, palliative care is not the primary focus of this work and therefore, results are not focused on capturing the subtle nuances of the field.

Frei and Kramer [26] have also made use of locally run GPT models to produce synthetic data. They ran a GPT NeoX language model using a few-shot prompting approach to generate structured German clinical text data. They subsequently used their data to train a NER model for classification tasks. However, this method relies on a large amount of computational power. Additionally, the clinical dataset is not focused on palliative care.

While the aforementioned studies have used LLMs, most have not focused on the creation of repeatable pipelines or frameworks to generate their information. While there is little research done into the use of LLMs for synthetic palliative data generation, investigations in the broader healthcare research space reveal that these models pose problems when trained on sensitive data, when they misunderstand their prompt input or when slight deviations in prompts produce dramatically different results [27]. To solve these issues, appropriate data, prompts and pipelines are needed for consistent and useful synthetic data.

LangChain is an open-source framework which aims to simplify LLM application development [28]. This framework has been used by Singh et al. [29] to investigate chatbot-type tools for patients. Additionally, Kapoor and Shetty [30] proposed a chatbot focused on general health using the framework. LangChain has also been used by Ke et al. [31] to create an LLM-based preoperative aid for healthcare professionals. However, as far as we are aware, LangChain has not yet been used for synthetic data generation in healthcare.

There is a need to investigate free-text note generation specifically for palliative nursing home settings. So far, research has focused on either hospital or mental health settings. Additionally, the majority of the models discussed in the aforementioned papers use real data in the training process. The use of GPT for synthetic palliative care data generation has not been thoroughly investigated as of yet. Furthermore, one-shot prompting, the generation of synthetic data from singular synthetic notes, and LangChain have not been fully investigated. We aim to address these research gaps in our approach.

3. Methods

In this section we outline our chosen models, prompt engineering techniques, generative process, and subsequent qualitative analysis procedure. Our methodology ensures that our system is scalable and flexible, allowing for adjustable inputs and extendable outputs. We also ensure that we only used easily accessible software, services and libraries, such that minimal financial or privacy-related barriers inhibit the recreation of our research or experimental results.

3.1. Choice of Model

To demonstrate the flexibility of LangChain and to compare model outputs, we performed the data generation process with both GPT-3.5 and GPT-4o models. GPT-3.5 is a transformer-based autoregressive language model; it is capable of producing output which is difficult to distinguish

from human text [32, 33]. This model has a knowledge base consisting of materials dated up to September, 2021 and has a 16,385-token context window [33]. GPT-4o is a more advanced, multi-modal transformer-based model which is designed to improve on some of the shortfalls of GPT-3.5 and GPT-4 [34, 33]. It is trained on data created prior to October, 2023 and has a 128,000-token context window [33]. At the time of writing the most recent GPT-4o model released is `gpt-4o-2024-05-13`; we compare this model with the `gpt-3.5-turbo-0125` GPT-3.5 model. Table 1 outlines the differences between these models [33]. In this study, we apply the same prompting and data "cleaning" process to the outputs of both models. Both models temperature settings are set at 1.1; we chose this parameter due to the use and validation of high temperature settings in past medical LLM research [35, 36]. Additionally, past research has also highlighted that statistically significant results and more creative outputs are more likely at temperatures over 1.0 [37, 38].

Table 1: GPT-3.5 vs GPT-4o Model Comparison

	`gpt-3.5-turbo-0125`	`gpt-4o-2024-05-13`
Context Window	16,385 Tokens	128,000 tokens
Max Output Tokens	4,096 tokens	4,096 tokens
Training Data	Up to Sep 2021	Up to Oct 2023

3.2. Process

We used Google Colab and the LangChain framework to implement our data generation process [39]. We used LangChain's `PromptTemplate` component to run similar prompts repeatedly [40]. In order to ensure a fluid flow between system "role" and data generation prompts, we used LangChain's `SequentialChain` component [41]. Each input note was passed to a function which could prompt the model using both a system "role" prompt and a

data generation prompt. The output of the model is subsequently "cleaned" and results are saved to CSV files. Each input note is used separately, and therefore there is an individual CSV file created for each associated output batch. Figure 1 outlines our note generation pipeline.

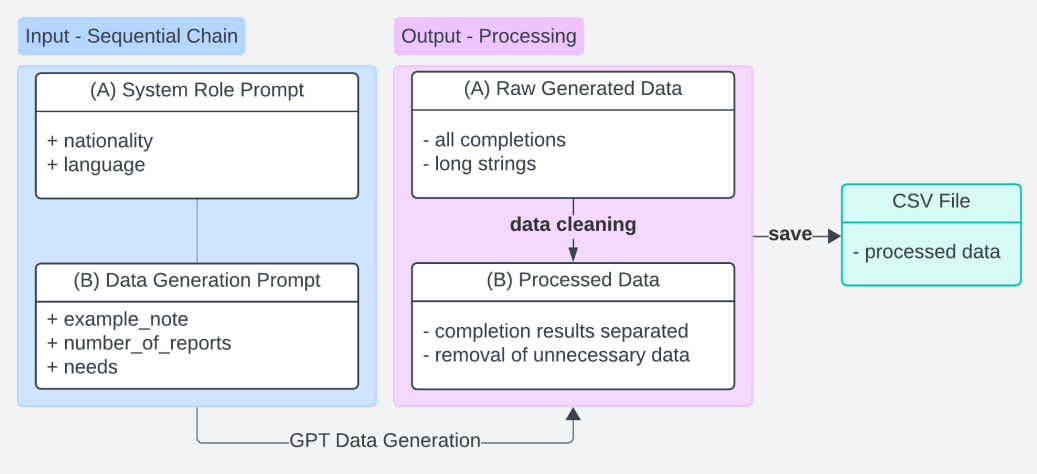


Figure 1: Note Generation Pipeline

3.3. Prompt Engineering

Prompt engineering involves the design of prompts which enable models to generate appropriate output across different domains [42]. These prompts can be optimized for specific tasks [42]. Prompt engineering is essential when interacting with LLMs to produce efficient, useful, and consistent generated output [43]. In the formulation of our prompts, we used many prompt engineering techniques. We also followed up-to-date guidelines provided for the medical field [43].

3.3.1. System Role Prompt

One common prompt engineering technique is to assign the model a "role" [43]. This "role" subsequently governs the model's data generation process and allows for more structured and consistent output. We decided to make

the model "role" flexible using LangChain; therefore, the associated prompt takes `nationality` and `language` as input variables. This flexibility allows our pipeline to be implemented by someone coming from a different cultural or geographical standpoint. Figure 2 describes this system "role" prompt.

```
system_role_prompt = {'message':
                      '''
                      You are a specialist in generating fictitious
data for natural language processing projects in healthcare.
                      You speak the language of a nurse in an
{nationality} nursing home. Namely, you speak {language}.
                      ''' ,
                      'inputs':["nationality", "language"]}
```

Figure 2: System Role Prompt

3.3.2. Data Generation Prompt

In line with prompt engineering techniques recommended for healthcare, we formulated the prompt input using expert knowledge [43]. Prompts were crafted in terms of directness and specificity using iterative testing; this approach allowed for focused responses. The topics included in the prompt were selected based on the content analysis of Dutch care notes using Latent Dirichlet Allocation (LDA) topic modeling. Additionally, the model is told not to generate overly-dramatic notes, as the majority of nursing home notes contain daily activities and mundane information. Figure 3 shows this data generation prompt.

GPT models have already proven themselves effective in resource-constrained situations [23]. Therefore, we decided to use a one-shot prompting approach to generate our notes, as recommended for healthcare research [43]. The model is provided with an example note to replicate. Input prompt notes were created by a qualified nurse. Each of these notes was designed to mimic a typical progress note which would be recorded for a resident during the daily documentation processes in a nursing home. The advantage of using

```
note_query_prompt = {'message':
                     '''
                     This is an example of a nurse note for a patient in a day:
                     "{example_note}"

                     Other reports may include: washing, dressing, brushing
                     teeth, getting ready for the day, getting ready for the
                     night, showering, cleaning dental prostheses, or assistance
                     after incontinence.
                     Other reports could include: what the client has or has not
                     eaten, what help is needed with eating (full help,
                     encouragement, adapted cutlery or cup), choking, keeping
                     hydration and nutrition lists.
                     Other reports could include: Organised activities, getting
                     visitors, browsing through a magazine, interacting with
                     fellow residents. Keep in mind that these are reports from
                     people in a nursing home, with severe disabilities, so
                     social interaction and activities are limited. Usually it
                     involves sociability, but not always.
                     Other reports may include, for example: oedema, pressure
                     ulcers, peeling, redness and itching of the skin. Nails that
                     are too long, blemishes.
                     Other reports could include, for example: care plan
                     discussions, minor medical complaints, family requests,
                     ordering medication.
                     Reports can be, for example, about: restlessness and
                     wandering at night, sleeping well, going to the toilet at
                     night, phoning, lying crookedly in bed.
                     Reports may include: agitation, restlessness, apathy,
                     confusion; usually the confusion is subtle, but sometimes
                     more intense.
                     Reports may include, for example: pain, tightness of breath,
                     nausea, diarrhoea, back pain, palliative care; usually the
                     complaints are subtle, but sometimes more severe.
                     Other reports can be about, for example: walking aids, the
                     wheelchair, falls, fall incidents, transfers, lifts.
                     Most reports are about everyday things, so not everything is
                     a serious incident.

                     Make up {number_of_reports} such reports for
                     {number_of_reports} residents with {needs} palliative care
                     needs. Return only the reports, with each report separated
                     by "***" and nothing else. Vary the sentence structure and
                     style.
                     ''' ,
                     'inputs':["example_note", "number_of_reports", "needs"]}
```

Figure 3: Data Generation Prompt

these notes is that they are realistic, but they are not from real residents; this ensures that generated data is free from privacy protection issues. Five notes described residents exhibiting signs of unmet palliative care needs. An additional five notes described residents not exhibiting signs of unmet palliative care needs. We labelled these notes "Unmet" and "Met" respectively. Table 2 outlines each note used in our study.

For the purposes of this study, we chose to execute 25 completions per input note and to generate 25 notes per completion. A completion is the model's generated output; it is a prediction of what is the most likely text to follow a given prompt [34]. Importantly, each completion is stateless when using the OpenAI API; every fresh completion does not remember the previous completion, leading to more diverse output and different writing styles within the notes. We decided that the model would generate 25 notes per completion in order to make the overall data generation process more efficient and diverse. However, there is no universally accepted standard in the number of completions or number of data points to use when prompting a model for synthetic data; every context requires a different choice[44, 45]. Prior work has shown that around 20-30 generations will lead to good diversity results, provided redundancy is not too large [18]. Our high number of completions was chosen with the knowledge that the models do not always generate appropriate output; we chose to use a high number of completions in order to select the best results from the model in later processing stages. Due to the flexibility of LangChain, the number of notes and completions can be adjusted to suit other use cases as future needs arise. Each of the 10 input notes was fed to the model individually, leading to 10 different prompting phases using the same prompt template. As we prompted the model to generate 25 notes for each of 25 completions, the total possible number of synthetic notes generated for each input note was 625. Therefore, with these 10 notes, our experimental setup had the potential to generate 6250 notes for each model used (GPT-3.5 and GPT-4o).

Table 2: 10 Input Notes

ID	Note	Needs
1	John was assisted with a shower this morning. He had his lunch in the canteen, he ate half a bowl of soup, half a chicken salad and a full bowl of ice cream along with 2 cups of tea. Johns' sister Mary came to see him today and he was in good form.	"Met"
2	Sarah had a GP appointment this morning and the GP was happy with her mobility progress. She has been advised to continue with her physiotherapy, but her pain relief has been reduced as her pain has been well controlled. The GP hopes that she will be able to walk without crutches within the next month.	"Met"
3	Michael showered independently this morning. It was his birthday today, so his family came to visit. His wife organised a sing song in the social room. Michael was in very good spirits and sang a few songs himself.	"Met"
4	Rachel was assisted with a wash this morning, her eyedrops were given, and moisturiser applied to her legs. She went for a walk in the garden in the afternoon and attended the painting activity in the social room. Rachel has advised that she needs new pyjamas and her family have contacted about same.	"Met"
5	Darragh has been started on an End of Life care plan. There was a family conference today including Darragh, his wife Eleanor, and his children. Goals of care have been discussed and his care plan has been updated. Darragh has no complaints of pain at present, however the GP has prescribed pain relief in case there is a need. Darragh is for regular turning to avoid breakdown of his skin and his skin is fully intact at present.	"Met"
6	Billy's sister has expressed her concern that his mobility is worsening, and he is becoming increasingly dependent on others for care.	"Unmet"
7	Stephen complained of pain 8/10 in his right heel this morning. His regular analgesia was given with little effect. Visually, his heel appears normal, and his skin is intact. However, Stephen is unable to weight bear on his right leg. Stephen has no PRN analgesia prescribed, and the GP has been contacted to review same.	"Unmet"
8	Andrea experienced nausea during lunch today. She was unable to eat and was given an anti-emetic for same. She returned to bed to lie down for an hour. Afterwards, she attempted to eat some toast and vomited immediately post. She has been unable to eat or drink since. She has been commenced on subcutaneous fluids and IM Zofran has been given. Awaiting GP review.	"Unmet"
9	Louise was assisted with a wash this morning. It was discovered that she had a 5cm lesion to her left buttock that contains slough and appears infected. A swab was taken and a wound chart commenced. Inadine and Leukomed (wound dressings) in situ at present. Awaiting review from Tissue Viability Nurse.	"Unmet"
10	Michelle was assisted with her breakfast this morning by one of the healthcare assistants. The healthcare assistant expressed her concern that Michelle's swallow seems to be worsening. A referral has been made for the Speech and Language Therapist to review. For now, we have placed Michelle on a level 5 diet and level 2 fluids until she is reviewed.	"Unmet"

3.4. Analysis

We chose a variety of quantitative and qualitative measures to evaluate our synthetic output. In this section, we describe details of these measures.

3.4.1. Qualitative Analysis

We performed a qualitative analysis based on human evaluation of the notes. Three subject matter experts, reviewed and blindly labelled samples

of synthetic notes, which the model had generated in response to either a "Met" or "Unmet" prompt. Approximately 10% of the GPT-3.5 notes and 10% of the GPT-4o notes were analysed. All GPT-3.5 notes were randomly shuffled and 600 notes were identified for review. The same shuffling process was repeated with the GPT-4o notes, and 650 notes were extracted. Both sets of notes contained an approximately equal amount of "Met" and "Unmet" notes. These notes were saved in Microsoft Excel files along with their corresponding labels. Separate Microsoft Excel files were prepared from which the labels were removed from the notes. These de-identified notes were subsequently sent to the three subject matter experts for blind relabelling. We also included the option to comment on each of the notes within each Excel file. Each of the three evaluators was not aware of the purpose of this labelling.

In order to effectively compare the results from our human evaluations and model outputs we calculated Fleiss' Kappa to assess the level of agreement between human results [46, 47]. We subsequently used majority voting to determine which human-generated labels to compare against the GPT labels. Furthermore we used Cohen's Kappa to determine the level of agreement between GPT-generated notes and human-generated labels.

3.4.2. Quantitative Analysis

BertScore is a contextual similarity metric created by Zhanget al. [48] which correlates better with human assessment than common text evaluation metrics. BERTScore aims to account for meaning in compositionally diverse text. BERTScore is calculated by computing the cosine similarity between two contextualized token embeddings. BERTScore reports precision, recall and F1 scores, and Zhang et al. recommend the use of the F1 score in evaluations. Using the `bert-score 0.3.13` Python library, each generated note was compared against its corresponding nurse-created example note [49]. As BERTScore is sensitive to punctuation, we removed all punctuation from our generated and comparison notes before computing BERTScore [50].

As outlined by Lavie and Agarwal [51], METEOR is a score used to represent lexical overlap. This metric is computed by matching words in an input sentence with those in an output sentence. METEOR has a high correlation level with human judgement. Like BERTScore, we compute METEOR metrics by comparing example input notes to generated output notes. We used NLTK's `meteor_score` module to calculate our results [52].

Additionally, we used the `TextBlob` Python library to calculate both sentiment and subjectivity as it has been used in other clinical literature [53]. `TextBlob`'s sentiment score is determined using a trained machine learning model, while the subjectivity measure is calculated by evaluating the "intensity" of each word in a sentence [54]. We compared the sentiment and subjectivity scores between the nurse-generated input, GPT-3.5 and GPT-4o notes using this library.

Due to its use in medical documentation research, we also used the Simple Measure Of Gobbledygook (SMOG) index to report on the readability of the generated notes [55, 56]. The `readability` Python library was used to calculate the SMOG index for the GPT-3.5 and GPT-4o notes [57]. Finally, we performed keyword analysis using the `wordcloud 1.9.3` Python library in order to gain an overview of the most common words found in the synthetic data [58].

We measured the time taken for both models to finish generating output using a completion speed test. We executed 5 completions per input note, with each completion generating 5 notes across 5 timing tests for each of the two models. Each of the tests was run on a Python 3 Google Compute Engine backend with 12.7GB of system RAM. Time was measured in seconds using Python's `time` library. The results of these tests were then averaged to determine which model took longer to generate output.

We also compared our chosen one-shot prompting style, as outlined in Figures 2 and 3, against a zero-shot approach. We used the BERTScore, METEOR, SMOG and sentiment analysis metrics outlined above to assess

the quality of our prompt with and without example-related prompt text given. We used 5 completions and 5 notes to compare both the original prompt and the zero-shot approach. The zero-shot system-role prompt remained the same. Figure 4 describes the data generation prompt used in the zero-shot comparison scenario. We used both GPT models in the evaluation of our prompt style and averaged the results across both models to create comparison metrics.

```
note_query_prompt = {
    'message:':
    '''
    Make up {number_of_reports} nurse notes for
    {number_of_reports} residents with {needs} palliative
    care needs. Return only the reports, with each report
    separated by "***" and nothing else. Vary the sentence
    structure and style.
    '''
}
```

Figure 4: Zero Shot Prompt

4. Results

In this section, we report on our experimental outputs and subsequent analysis. As outlined in our methodology, we used "role" formulation, direct language, example input, and iteration prompt engineering techniques, along with the OpenAI API and the LangChain framework, to generate our synthetic data. All qualitative and quantitative results described below as percentages are reported to two decimal places and all fractional results are reported to four decimal places.

4.1. Model Outputs

There were a total of 12036 notes created from a combination of both the GPT-3.5 and GPT-4o processes. Table 3 outlines detailed numerical information about the generated outputs, including the number of notes after manual error removal. Overall, the majority of notes were preserved during post-processing. While each model was capable of producing 6250 notes using our pipeline, this number is not fully reflected in the dataset due to minor processing issues. These data processing issues involved the models producing output which was not properly delineated; each note was supposed to be separated by "***", but some model outputs did not include appropriate use of this formatting instruction. As a result, some model outputs were saved as extremely long notes; the majority of these were caught in initial processing stages after the model completion and before data saving. However, the few noisy entries that made it through to the final data were manually removed after all model completions were saved.

4.2. Qualitative Analysis

Our human evaluation results, based on nurse-labelling of the generated notes, are outlined below.

4.2.1. GPT-3.5 Results

3 of the 600 GPT-3.5 notes (0.5%) were deemed to be "problematic". "Problematic" within this context refers to notes that were thought to be too difficult to classify by the subject matter experts due to file formatting issues, language ambiguity, or misspellings. Problematic notes were subsequently removed from this analysis, leaving 597 notes for use.

The Fleiss' Kappa score obtained when all three human-raters were compared was 0.58 for the 597 usable GPT-3.5 notes, indicating high moderate agreement between all three raters. 487 of the 597 notes (81.57%) were labelled similarly by both the model and the expert. The Cohen's Kappa coefficient calculated for these notes was approximately 0.6331, indicating

Table 3: Output Note Categorization

	GPT-3.5 Before Manual Processing	GPT-3.5 After Manual Processing	GPT-4o Before Manual Processing	GPT-4o After Manual Processing
Note 1 (Met)	594	592	628	628
Note 2 (Met)	575	573	610	610
Note 3 (Met)	577	575	637	637
Note 4 (Met)	570	569	636	636
Note 5 (Met)	588	588	587	587
Note 6 (Unmet)	579	578	610	610
Note 7 (Unmet)	571	569	643	643
Note 8 (Unmet)	553	552	642	642
Note 9 (Unmet)	598	597	640	640
Note 10 (Unmet)	578	577	620	620
Total Unmet	2879	2873	3155	3155
Total Met	2904	2897	3098	3098
Total	5783	5770	6253	6253

substantial agreement between the subject matter experts and the GPT-3.5 model. The F1 score calculated based on these notes was 83.78%, indicating that the model was capable of generating valid synthetic nursing home resident notes which could be used for further analysis and research.

Figure 5 describes a confusion matrix outlining a detailed breakdown of the classification outputs. From this matrix, one can see that the experts categorised the majority of notes correctly.

4.2.2. GPT-4o Results

1 of the 650 GPT-4o notes (0.15%) presented to the experts was deemed to be "problematic". Similarly to the GPT-3.5 analysis, this note was excluded from subsequent assessments, leaving 649 notes for further use.

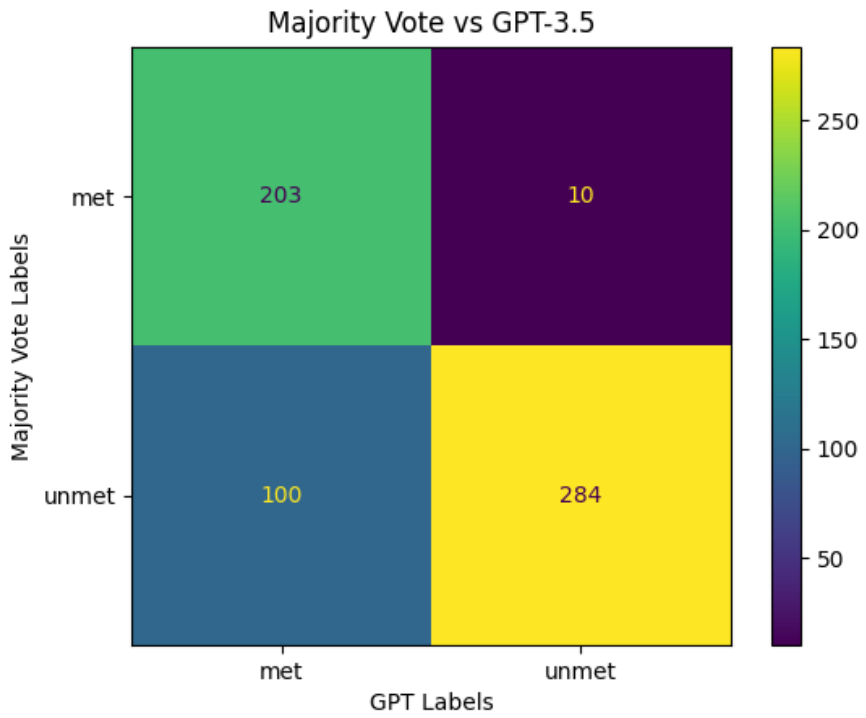


Figure 5: GPT-3.5 vs Human Experts Confusion Matrix

The Fleiss' Kappa score obtained when comparing our human-raters to the model was 0.299, indicating fair agreement between all three raters. 408 of the 649 notes (62.87%) were labelled similarly by both the model and the expert. The Cohen's Kappa coefficient calculated for these notes was approximately 0.2674, indicating fair agreement between the subject matter experts and the GPT-4o model. The F1 score calculated based on these notes was 68.98%, indicating that the model performed poorly in comparison to the GPT-3.5 model.

Figure 6 describes a confusion matrix comprising the expert and GPT-4o labels. From this matrix, one can see that there is a significant disimprovement in the results of the GPT-4o model in comparison to the GPT-3.5 model.

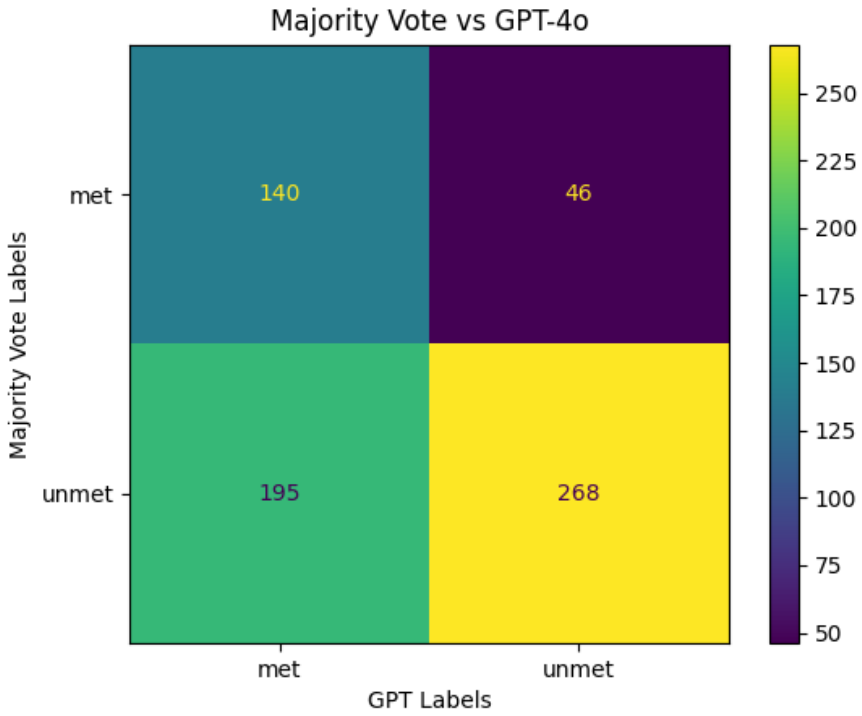


Figure 6: GPT-4o vs Human Experts Confusion Matrix

4.2.3. Overall Qualitative Results

Informal feedback from the experts indicated that the majority of notes were "quite good". While these notes do not feature many clinical terms, the relaxed tone and informal recording of resident conditions and care seem to be in keeping with writing styles from other nursing home note analyses [59, 60]. The GPT-3.5 notes seem to be more faithful to human opinion than the GPT-4o notes when compared using the F1 and Cohen's Kappa coefficient results of our test. Additionally, there was more inter-rater agreement in the GPT-3.5 labelling evaluation when compared against the GPT-4o evaluation. These findings indicate that, when using human-based assessments, the GPT-3.5 model is more suited to nursing home resident synthetic data creation tasks when compared against the GPT-4o model.

4.3. Quantitative Analysis

We performed a quantitative analysis of all notes for both the GPT-3.5 and GPT-4o implementations using the metrics outlined in our methodology.

4.3.1. BERTScore - Contextual Similarity

Table 4 outlines the BERTScore F1 results for each GPT model. While somewhat contextually similar, we can see that the models produced outputs that exhibit appropriate deviation from the input content. Our findings show that the generated notes were sufficiently diverse enough to create appropriately heterogeneous datasets related to the context.

Table 4: BERTScore F1 Evaluation

	GPT-3.5 (%)	GPT-4o (%)
Note 1 (Met)	51.30	51.95
Note 2 (Met)	47.65	48.2
Note 3 (Met)	48.91	50.33
Note 4 (Met)	54.16	53.53
Note 5 (Met)	55.91	48.81
Note 6 (Unmet)	49.97	45.39
Note 7 (Unmet)	51.18	52.5
Note 8 (Unmet)	48.99	50.40
Note 9 (Unmet)	47.24	49.90
Note 10 (Unmet)	47.34	48.40
Average Unmet	48.94	49.32
Average Met	51.59	50.56
Overall Average	50.27	49.94

4.3.2. METEOR - Lexical Overlap

METEOR scores for GPT-3.5 and GPT-4o synthetic data are outlined in Table 5. These results indicate that the GPT models used a wide vocabulary to generate varied results. This variance shows that the models did

not "overfit" to the one-shot input notes, but remained appropriately varied throughout the prompting process.

Table 5: METEOR Evaluation Reported Using Decimal Notation

	GPT-3.5	GPT-4o
Note 1 (Met)	0.1851	0.1753
Note 2 (Met)	0.1388	0.1215
Note 3 (Met)	0.1204	0.1206
Note 4 (Met)	0.2024	0.1904
Note 5 (Met)	0.2728	0.1343
Note 6 (Unmet)	0.0996	0.0712
Note 7 (Unmet)	0.1371	0.1379
Note 8 (Unmet)	0.1178	0.1192
Note 9 (Unmet)	0.0956	0.1296
Note 10 (Unmet)	0.1485	0.1524
Average Unmet	0.1197	0.1221
Average Met	0.1839	0.1484
Overall Average	0.1518	0.1352

4.3.3. Lexicon-Based Sentiment Analysis

Overall "Met" notes tended to be more positive than "Unmet" notes. Subjectivity results were fairly low overall, indicating that such notes are not generally subjective. Average results are reported in Table 6. From these results, we can see that both sets of model output resulted in similar sentiment and subjectivity metrics. The model results are also quite similar to those from the input nurse notes, indicating that the correct tone is reflected in these notes.

4.3.4. Readability

Overall, GPT-3.5 notes resulted in a SMOG index of 10.29, indicating a 10th grade reading level. GPT-4o results revealed a similar reading level with

Table 6: Sentiment and Subjectivity Reported Using Decimal Notation

	GPT-3.5	GPT-4o	Nurse Notes
Sentiment "Met" Notes	0.1798	0.1164	0.1422
Sentiment "Unmet" Notes	0.0416	0.0599	0.0733
Subjectivity "Met" Notes	0.3913	0.3922	0.3244
Subjectivity "Unmet" Notes	0.2936	0.3536	0.3650
Nurse-Model "Met" Sentiment Difference	0.0376	0.0258	-
Nurse-Model "Unmet" Sentiment Difference	0.1065	0.0430	-
Nurse-Model "Met" Subjectivity Difference	0.0669	0.0677	-
Nurse-Model "Unmet" Subjectivity Difference	0.0263	0.0272	-

a SMOG index of 10.69. Although the recommended reading level for medical text is 6th grade, the results from these models are in keeping with the higher grade levels of between 9 and 15 observed in freely available medical notes [61, 55]. Therefore, while simpler notes are recommended, GPT notes may be more faithful to the types of notes used in practice.

4.3.5. Keyword Analysis

The top words for each model and note category ("Met" or "Unmet") are reported in Table 7. We removed the word "Mr" from the analysis after initial inspection as "Mr" is a shortening of "mister" and is not a word of interest per se.

Positive action words such as "participated" and "enjoyed" are present across both GPT-3.5 and GPT-4o "Met" notes. "Unmet" GPT-3.5 notes seem to focus on negative words such as "discomfort", whereas GPT-4o "Unmet" notes seemed to remain fairly neutral. Figure 7 describes the overall distribution of words across all generated notes. This figure reveals a heavy emphasis on temporal words, such as "today", "morning", and "afternoon".

Table 7: Top 10 Keywords for Each Model and Note Category

GPT-3.5 "Met"	GPT-3.5 "Unmet"	GPT-4o "Met"	GPT-4o "Unmet"
today	palliative care	morning	today
care plan	need	today	bed
visit	confusion	afternoon	morning
enjoyed	care team	lunch	required
afternoon	review	day	day
morning	agitation	bed	afternoon
day	restlessness	breakfast	noted
experienced	discomfort	enjoyed	night
participated	complained	participated	staff
requested	exhibited signs	seemed	wheelchair

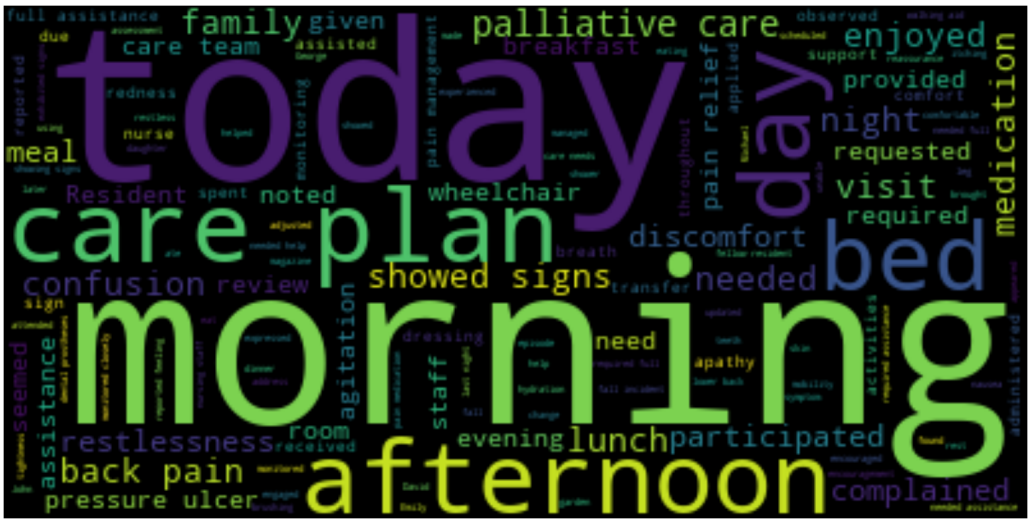


Figure 7: All Notes Word Cloud

4.3.6. Timing Tests

The results of our timing tests are outlined in Table 8. Clearly from this table, GPT-3.5 is a lot faster at completing the note generation task than

GPT-4o, indicating that the GPT-3.5 model is more efficient than its more recent counterpart.

Table 8: Model Completion Time

Run Number	GPT-3.5 (seconds)	GPT-4o (seconds)
Run 1	223.39	1212.10
Run 2	239.47	1085.89
Run 3	287.74	997.09
Run 4	217.71	1118.24
Run 5	211.34	1117.94
Average Time	235.93	1106.25

4.3.7. Prompt Comparisons

The results of our prompt style assessments are outlined in Table 9. From this table, it is clear that our main one-shot approach outperforms the zero-shot approach across metrics. The SMOG index results indicate that the one-shot approach results in notes of a more manageable reading level than the zero-shot approach. The BERTScore results indicate that the one-shot approach results in notes that are more similar to our nurse-generated input notes. The METEOR score also conveys the slightly better performance of the one-shot approach. There are also greater differences in subjectivity and sentiment across both met and unmet notes for zero-shot results in comparison to one-shot results, indicating that the one-shot approach results in notes that are more similar to the input notes. While the majority of results demonstrate only small improvements across metrics, these findings indicate that the one-shot guidance results in notes that are more appropriate to the context.

Additionally, the zero-shot approach resulted in many notes being incorrectly generated without proper delineation. The zero-shot prompts resulted

in a total of 119 notes. The one-shot prompts resulted in a total of 480 notes. Given that the total number of notes expected given our 5-note 5-completion test parameters was 500, these results indicate that the one-shot approach generated more appropriate output in this scenario.

Table 9: Prompt Style Comparison

Metric	Average Zero-Shot	Average One-Shot
SMOG	12.27	11.67
BERTScore	0.44	0.54
METEOR	0.07	0.16
Sentiment "Met" Difference	0.11	0.05
Sentiment "Unmet" Difference	0.04	0.03
Subjectivity "Met" Difference	0.1	0.09
Subjectivity "Unmet" Difference	0.13	0.025

5. Discussion

This paper presents a template LangChain and subsequent validation method for generating synthetic data for nursing home palliative care research. This template can be easily modified to include different topics as needed. As a result of our implementation, we were able to generate over 12,000 notes using just 10 nurse-provided examples. We investigated the use of such a template to produce notes which are reflective of human-generated notes. These notes were evaluated using a variety of qualitative and quantitative measures. We assessed the feasibility of our approach using these measures. We also compared GPT-3.5 and GPT-4o models using these assessments.

Both the GPT-3.5 and GPT-4o models exhibit similar results when compared using our quantitative assessments. Keyword analysis revealed that both models use temporal words. As described in Section 4.3.1, we used

BERTScore to compute contextual similarity; we found both models capable of generating appropriately contextually-diverse data. METEOR was used to examine lexical overlap, as outlined in Section 4.3.2; there is suitable variance in the generated data. Section 4.3.4 outlines the result of our SMOG readability assessments; our notes were of similar reading level to other human-written nurse notes. Additionally, lexicon-based sentiment analysis, which is described in Section 4.3.3, revealed that our generated content reflects similar sentiment to our nurse-generated input notes.

While the models convey similar results when observed using quantitative assessments, our qualitative evaluation reveals that GPT-3.5 notes are more in-keeping with expert opinion than GPT-4o notes. We conducted a blind relabelling test to find that the GPT-3.5 output is more reflective of human generated notes than the GPT-4o output. Given the importance of the qualitative nature of nurse notes, we posit that GPT-3.5 is more suitable for the note generation task. Furthermore, upon assessing the amount of time taken to complete the note generation task, GPT-3.5 outperformed GPT-4o by a significant margin, demonstrating the efficiency of the former model in comparison to the latter.

Furthermore, GPT-4o's ability to handle lengthier text with increased complexity may be disadvantageous in simple and concise note generation contexts. Through these results, we show that more recent models do not necessarily outperform older models across all metrics or evaluations. This finding implies that, when generating synthetic data, multiple models should be considered, not just the most recent versions.

This study is not without limitations. All of the nurse notes used as example input during testing were generated by one nurse. The nurse expert has been trained and works in the Irish healthcare system only and has no experience of other healthcare systems. These generated notes may not be relevant outside an Irish context. Future experiments could include notes from a wider variety of health professionals to further confirm the feasibility

of our approach.

While we attempted to evaluate our notes using both qualitative and quantitative analysis, there exists no standard evaluation for synthetically generated text [62]. While human evaluation is considered the gold standard, there are no specific criteria outlined for such assessments [62]. This lack of standardisation extends to quantitative evaluation scenarios also. The context of this work makes it difficult to compare our findings to other synthetic data studies. Additionally, in life-and-death settings, higher standards are required, rather than relying on arbitrary assessment techniques.

Our human evaluation is carried out on a 10% sample of randomly shuffled generated notes due to time and resource constraints. While such assessment is useful in determining a general overview of data quality, this approach does not guarantee that our results are representative of the entire data domain. Additionally, one of our evaluators generated the input notes for the experiment. While this evaluator was blind to the purpose of the evaluation and was not involved in the design of the experimental procedure, the use of the same nurse in both generation and evaluation processes could have potentially introduced bias into our study.

Furthermore, in an effort to discretize the palliative care referral landscape, we classify notes into either "Met" or "Unmet" data points. In palliative care practice, it is more common to perceive such data as indicative or not indicative of signs of unmet palliative care needs. Therefore, our human evaluation may have resulted in an assessment skewed towards the identification of unmet palliative care needs. Furthermore, discretizing such a complex field may potentially hinder future use of such data. A more structured approach to labelling notes with pre-defined criteria may improve results.

While our synthetic data usage aims to ensure that sensitive information from real patients is not revealed, we cannot guarantee that the foundation models being used do not leak training data. LLMs are trained on a wide

variety of information from diverse sources; we do not currently have access to the exact data that these models were trained on. Additionally, there is no definitive way to check for the leakage of sensitive information from these models due to the lack of knowledge surrounding the initial foundation model training data. Future work could focus on using open-source models with more clarity surrounding leakage risks to ensure safer synthetic dataset generation processes.

Our results demonstrate the feasibility of using LangChain, the OpenAI API, and one-shot prompting to generate synthetic data for future research. However, a further blind study is needed involving a panel of experts in order to validate this data for future use. If such synthetic data is valid, researchers can use it in preliminary investigation stages before trying to access real-world notes. These notes could be used to test potential algorithms before real resident data is accessed and used by researchers. Further work could compare the performance of machine learning models trained on synthetically generated data in comparison to the same models trained on real resident data when used on unseen data. This future assessment could inform future research on the functionality of synthetic data.

6. Conclusion

This pilot study outlines a novel synthetic data generation pipeline to investigate the feasibility of using GPT models to generate nursing home resident data for palliative care research. We also compare results between these models. Our findings indicate that such approaches to artificial data creation can result in synthetic notes which may be reflective of human-generated data. We also find that GPT-3.5 may be more suited to such data generation tasks when compared against GPT-4o using human evaluation. In this paper, we discuss our methodology, analysis, and future directions. We also outline potential limitations associated with our work. Such data has the potential to be used by future researchers to accelerate palliative care

research and to preserve patient privacy.

Appendix A. Supplementary Data

The code used and datasets generated during this research can be found on GitHub at `https://github.com/isabel-ronan/SyntheticNotes`.

CRediT Authorship Contribution Statement

Isabel Ronan: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Conceptualization, Validation, Formal Analysis, Data Curation. **Patrice Crowley**: Writing – review & editing, Investigation, Conceptualization, Validation. **Eva Rombouts**: Writing – review & editing, Conceptualization, Validation. **Nicola Cornally**: Writing – review & editing, Supervision, Conceptualization. **Mohamad M Saab**: Writing – review & editing, Supervision, Conceptualization, Validation. **David Murphy**: Writing – review & editing, Supervision, Conceptualization. **Sabin Tabirca**: Writing – review & editing, Supervision, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Sources

This publication has emanated from research conducted with the financial support of Taighde Éireann – Research Ireland under Grant number 18/CRT/6222. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Patrice Crowley is in receipt of a PhD scholarship from the Catherine McAuley School of Nursing and Midwifery, University College Cork, Cork, Ireland.

References

[1] J. L. Powell, The power of global aging 35 (1) 1–14. doi:10.1007/s12126-010-9051-6.
URL https://doi.org/10.1007/s12126-010-9051-6

[2] A. J. De Veer, A. Kerkstra, Feeling at home in nursing homes 35 (3) 427–434, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1365-2648.2001.01858.x. doi:10.1046/j.1365-2648.2001.01858.x.
URL https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2648.2001.01858.x

[3] L. Schenk, R. Meyer, A. Behr, A. Kuhlmey, M. Holzhausen, Quality of life in nursing homes: results of a qualitative resident survey 22 (10) 2929–2938. doi:10.1007/s11136-013-0400-2.
URL https://doi.org/10.1007/s11136-013-0400-2

[4] I. Hjaltadóttir, M. Gústafsdóttir, Quality of life in nursing homes: perception of physically frail elderly residents 21 (1) 48–55. doi:10.1111/j.1471-6712.2007.00434.x.
URL https://pubmed.ncbi.nlm.nih.gov/17428214/

[5] S. Hall, S. Longhurst, I. Higginson, Living and dying with dignity: a qualitative study of the views of older people in nursing homes 38 (4) 411–416. doi:10.1093/ageing/afp069.
URL https://doi.org/10.1093/ageing/afp069

[6] Topics in palliative care.
URL http://ndl.ethernet.edu.et/bitstream/123456789/45753/1/3228.pdf

[7] B. Wallerstedt, E. Benzein, K. Schildmeijer, A. Sandgren, What is palliative care? perceptions of healthcare professionals 33 (1) 77–84. doi:10.1111/scs.12603.

[8] W. Van Mechelen, B. Aertgeerts, K. De Ceulaer, B. Thoonsen, M. Vermandere, F. Warmenhoven, E. Van Rijswijk, J. De Lepeleire, Defining the palliative care patient: A systematic review 27 (3) 197–208, publisher: SAGE Publications Ltd STM. doi:10.1177/0269216311435268.
URL https://doi.org/10.1177/0269216311435268

[9] N. M. Cimino, M. L. McPherson, Evaluating the impact of palliative or hospice care provided in nursing homes 40 (10) 10–14. doi:10.3928/00989134-20140909-01.

[10] R. Schweighart, J. L. O'Sullivan, M. Klemmt, A. Teti, S. Neuderth, Wishes and needs of nursing home residents: A scoping review 10 (5) 854. doi:10.3390/healthcare10050854.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9140474/

[11] J. M. Helm, A. M. Swiergosz, H. S. Haeberle, J. M. Karnuta, J. L. Schaffer, V. E. Krebs, A. I. Spitzer, P. N. Ramkumar, Machine learning and artificial intelligence: Definitions, applications, and future directions 13 (1) 69–76. doi:10.1007/s12178-020-09600-8.
URL https://doi.org/10.1007/s12178-020-09600-8

[12] V. Reddy, A. Nafees, S. Raman, Recent advances in artificial intelligence applications for supportive and palliative care in cancer patients 17 (2) 125–134. doi:10.1097/SPC.0000000000000645.

[13] G. Hansebo, M. Kihlgren, G. Ljunggren, Review of nursing documentation in nursing home wards — changes after intervention for individualized care 29 (6) 1462–1473, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1365-2648.1999.01034.x. doi:10.1046/j.1365-2648.1999.01034.x.

URL https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2648.1999.01034.x

[14] P. Voutilainen, A. Isola, S. Muurinen, Nursing documentation in nursing homes – state-of-the-art and implications for quality improvement 18 (1) 72–81, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1471-6712.2004.00265.x. doi:10.1111/j.1471-6712.2004.00265.x.
URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1471-6712.2004.00265.x

[15] A. Hendriks, C. Hacking, H. Verbeek, S. Aarts, Data science techniques to gain novel insights into quality of care: a scoping review of long-term care for older adults 2 (2) 67–85, number: 2 Publisher: Open Exploration. doi:10.37349/edht.2024.00012.
URL https://www.explorationpub.com/Journals/edht/Article/101112

[16] A. Figueira, B. Vaz, Survey on synthetic data generation, evaluation methods and GANs 10 (15) 2733, number: 15 Publisher: Multidisciplinary Digital Publishing Institute. doi:10.3390/math10152733.
URL https://www.mdpi.com/2227-7390/10/15/2733

[17] W. Hahn, K. Schütte, K. Schultz, O. Wolkenhauer, M. Sedlmayr, U. Schuler, M. Eichler, S. Bej, M. Wolfien, Contribution of synthetic data generation towards an improved patient stratification in palliative care 12 (8) 1278, number: 8 Publisher: Multidisciplinary Digital Publishing Institute. doi:10.3390/jpm12081278.
URL https://www.mdpi.com/2075-4426/12/8/1278

[18] R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. K. Williamson, F. Mahmood, Synthetic data in machine learning for medicine and healthcare 5 (6) 493–497, publisher: Nature Publishing Group. doi:10.1038/s41551-021-

00751-8.
URL https://www.nature.com/articles/s41551-021-00751-8

[19] H. Murtaza, M. Ahmed, N. F. Khan, G. Murtaza, S. Zafar, A. Bano, Synthetic data generation: State of the art in health care domain 48 100546. doi:10.1016/j.cosrev.2023.100546.
URL https://www.sciencedirect.com/science/article/pii/S1574013723000138

[20] D. Rankin, M. Black, R. Bond, J. Wallace, M. Mulvenna, G. Epelde, Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for data sharing 8 (7) e18910, company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics Publisher: JMIR Publications Inc., Toronto, Canada. doi:10.2196/18910.
URL https://medinform.jmir.org/2020/7/e18910

[21] E. Begoli, K. Brown, S. Srinivas, S. Tamang, SynthNotes: A generator framework for high-volume, high-fidelity synthetic mental health notes, in: 2018 IEEE International Conference on Big Data (Big Data), pp. 951–958. doi:10.1109/BigData.2018.8621981.
URL https://ieeexplore.ieee.org/document/8621981

[22] C. A. Libbi, J. Trienes, D. Trieschnigg, C. Seifert, Generating synthetic training data for supervised de-identification of electronic health records 13 (5) 136, number: 5 Publisher: Multidisciplinary Digital Publishing Institute. doi:10.3390/fi13050136.
URL https://www.mdpi.com/1999-5903/13/5/136

[23] A. Amin-Nejad, J. Ive, S. Velupillai, Exploring transformer text generation for medical dataset augmentation, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara,

B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, pp. 4699–4708.
URL https://aclanthology.org/2020.lrec-1.578

[24] F. Sufi, Addressing data scarcity in the medical domain: A GPT-based approach for synthetic data generation and feature extraction 15 (5) 264, number: 5 Publisher: Multidisciplinary Digital Publishing Institute. doi:10.3390/info15050264.
URL https://www.mdpi.com/2078-2489/15/5/264

[25] H. Šuvalov, M. Lepson, V. Kukk, M. Malk, N. Ilves, H.-A. Kuulmets, R. Kolde, Using synthetic health care data to leverage large language models for named entity recognition: Development and validation study 27 (1) e66279, company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada. doi:10.2196/66279.
URL https://www.jmir.org/2025/1/e66279

[26] J. Frei, F. Kramer, Annotated dataset creation through general purpose language models for non-english medical NLP. arXiv:2208.14493 [cs], doi:10.48550/arXiv.2208.14493.
URL http://arxiv.org/abs/2208.14493

[27] K. He, R. Mao, Q. Lin, Y. Ruan, X. Lan, M. Feng, E. Cambria, A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. arXiv:2310.05694 [cs], doi:10.48550/arXiv.2310.05694.
URL http://arxiv.org/abs/2310.05694

[28] H. Chase, LangChain, original-date: 2022-10-17T02:58:36Z.
URL https://github.com/langchain-ai/langchain

[29] A. Singh, A. Ehtesham, S. Mahmud, J.-H. Kim, Revolutionizing mental health care through LangChain: A journey with a large language model, in: 2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC), pp. 0073–0078. doi:10.1109/CCWC60891.2024.10427865.
URL https://ieeexplore.ieee.org/abstract/document/10427865

[30] A. Kapoor, S. D. Shetty, Enhancing healthcare information accessibility through a generative medical chatbot, in: 2024 International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications (ICETCS), IEEE, pp. 1–4. doi:10.1109/ICETCS61022.2024.10543529.
URL https://ieeexplore.ieee.org/document/10543529/

[31] Y. H. Ke, L. Jin, K. Elangovan, H. R. Abdullah, N. Liu, A. T. H. Sia, C. R. Soh, J. Y. M. Tung, J. C. L. Ong, D. S. W. Ting, Development and testing of retrieval augmented generation in large language models - a case study report.

[32] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners. arXiv:2005.14165 [cs], doi:10.48550/arXiv.2005.14165.
URL http://arxiv.org/abs/2005.14165

[33] OpenAI platform.
URL https://platform.openai.com

[34] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila,

I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. d. A. B. Peres, M. Petrov, H. P. d. O. Pinto, Michael, Pokorny, M. Pokrass, V. H. Pong, T. Powell,

A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. J. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, B. Zoph, GPT-4 technical report. arXiv:2303.08774 [cs], doi:10.48550/arXiv.2303.08774.
URL `http://arxiv.org/abs/2303.08774`

[35] M. Tornqvist, J.-D. Zucker, T. Fauvel, N. Lambert, M. Berthelot, A. Movschin, A text-to-tabular approach to generate synthetic patient data using LLMs. arXiv:2412.05153 [cs], doi:10.48550/arXiv.2412.05153.
URL `http://arxiv.org/abs/2412.05153`

[36] D. Smolyak, A. Welivita, M. V. Bjarnadóttir, R. Agarwal, Improving equity in health modeling with GPT4-turbo generated synthetic data: A comparative study. arXiv:2412.16335 [cs], doi:10.48550/arXiv.2412.16335.
URL `http://arxiv.org/abs/2412.16335`

[37] L. Li, L. Sleem, N. Gentile, G. Nichil, R. State, Exploring the impact of temperature on large language models:hot or cold? arXiv:2506.07295 [cs], doi:10.48550/arXiv.2506.07295.
URL `http://arxiv.org/abs/2506.07295`

[38] M. Renze, E. Guven, The effect of sampling temperature on problem solving in large language models, pp. 7346–7356. arXiv:2402.05201 [cs], doi:10.18653/v1/2024.findings-emnlp.432.
URL http://arxiv.org/abs/2402.05201

[39] colab.google.
URL https://colab.google/

[40] langchain_core.prompts.prompt.PromptTemplate — LangChain 0.2.7.
URL https://api.python.langchain.com/en/latest/prompts/langchain_core.prompts.prompt.PromptTemplate.html

[41] langchain.chains.sequential.SequentialChain — LangChain 0.2.7.
URL https://api.python.langchain.com/en/latest/chains/langchain.chains.sequential.SequentialChain.html

[42] J. Wang, E. Shi, S. Yu, Z. Wu, C. Ma, H. Dai, Q. Yang, Y. Kang, J. Wu, H. Hu, C. Yue, H. Zhang, Y. Liu, Y. Pan, Z. Liu, L. Sun, X. Li, B. Ge, X. Jiang, D. Zhu, Y. Yuan, D. Shen, T. Liu, S. Zhang, Prompt engineering for healthcare: Methodologies and applications. arXiv:2304.14670 [cs], doi:10.48550/arXiv.2304.14670.
URL http://arxiv.org/abs/2304.14670

[43] B. Meskó, Prompt engineering as an important emerging skill for medical professionals: Tutorial 25 (1) e50638, company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada. doi:10.2196/50638.
URL https://www.jmir.org/2023/1/e50638

[44] H. Chen, A. Waheed, X. Li, Y. Wang, J. Wang, B. Raj, M. I. Abdin, On the diversity of synthetic data and its impact on training large language

models. arXiv:2410.15226 [cs], doi:10.48550/arXiv.2410.15226.
URL http://arxiv.org/abs/2410.15226

[45] Y. Zhu, H. Zhong, Q. Lin, H. Wei, X. Sun, Z. Yu, M. Liu, Z. Zheng, L. Chen, What matters in LLM-generated data: Diversity and its effect on model fine-tuning. arXiv:2506.19262 [cs], doi:10.48550/arXiv.2506.19262.
URL http://arxiv.org/abs/2506.19262

[46] T. R. Nichols, P. M. Wisner, G. Cripe, L. Gulabchand, Putting the kappa statistic to use 13 (3) 57–61, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qaj.481. doi:10.1002/qaj.481.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/qaj.481

[47] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data 33 (1) 159–174, publisher: International Biometric Society. doi:10.2307/2529310.
URL https://www.jstor.org/stable/2529310

[48] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT. arXiv:1904.09675 [cs], doi:10.48550/arXiv.1904.09675.
URL http://arxiv.org/abs/1904.09675

[49] Artzi, Yoav, T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, bert-score: PyTorch implementation of BERT score.
URL https://github.com/Tiiiger/bert_score

[50] P. Jwalapuram, Pulling out all the full stops: Punctuation sensitivity in neural machine translation and evaluation, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computa-

tional Linguistics: ACL 2023, Association for Computational Linguistics, pp. 6116–6130. doi:10.18653/v1/2023.findings-acl.381.
URL https://aclanthology.org/2023.findings-acl.381

[51] A. Lavie, A. Agarwal, Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments, in: Proceedings of the Second Workshop on Statistical Machine Translation - StatMT '07, Association for Computational Linguistics, pp. 228–231. doi:10.3115/1626355.1626389.
URL http://portal.acm.org/citation.cfm?doid=1626355.1626389

[52] NLTK :: nltk.translate.meteor_score module.
URL https://www.nltk.org/api/nltk.translate.meteor_score.html

[53] K. Denecke, D. Reichenpfader, Sentiment analysis of clinical narratives: A scoping review 140 104336. doi:10.1016/j.jbi.2023.104336.
URL https://www.sciencedirect.com/science/article/pii/S1532046423000576

[54] S. Loria, sloria/TextBlob, original-date: 2013-06-30T18:29:18Z.
URL https://github.com/sloria/TextBlob

[55] A. S. Hedman, Using the SMOG formula to revise a health-related document 39 (1) 61–64. doi:10.1080/19325037.2008.10599016.

[56] P. Fitzsimmons, B. Michael, J. Hulley, G. Scott, A readability assessment of online parkinson's disease information 40 (4) 292–296. doi:10.4997/JRCPE.2010.401.
URL http://www.rcpe.ac.uk/journal/issue/journal_40_4/fitzsimmons.pdf

[57] A. v. Cranenburgh, readability: Measure the readability of a given text using surface characteristics.
URL https://github.com/andreasvc/readability/

[58] A. Mueller, wordcloud: A little word cloud generator.
URL https://github.com/amueller/word_cloud

[59] I. Gunhardsson, A. Svensson, C. Berterö, Documentation in palliative care: Nursing documentation in a palliative care unit—a pilot study 25 (1) 45–51, publisher: SAGE Publications Inc. doi:10.1177/1049909107307381.
URL https://doi.org/10.1177/1049909107307381

[60] M. C. Broderick, A. Coffey, Person-centred care in nursing documentation 8 (4) 309–318, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/opn.12012. doi:10.1111/opn.12012.
URL https://onlinelibrary.wiley.com/doi/abs/10.1111/opn.12012

[61] W.-C. Su, K. Dufendach, D. T. Wu, Assessing the readability of freely available ICU notes 2019 696–703.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6568110/

[62] A. Celikyilmaz, E. Clark, J. Gao, Evaluation of text generation: A survey. arXiv:2006.14799 [cs], doi:10.48550/arXiv.2006.14799.
URL http://arxiv.org/abs/2006.14799