

Master Thesis

Chats, Agents and Lyrics

Alyssa MacGregor-Hastie

*a thesis submitted in partial fulfilment of the requirements
for the degree of*

MA Linguistics

(Text Mining)

Vrije Universiteit Amsterdam

Computational Lexicology and Terminology Lab
Department of Language and Communication
Faculty of Humanities



Supervised by: Isa Maks
2nd reader: Ilia Markov

Submitted: March 26, 2024

Abstract

This thesis focuses on evaluating ChatGPT's ability to perform prompt-guided topic extraction on song lyrics. Prompt-guided topic extraction is the task of guiding ChatGPT to assign topic labels to a given text - in our case, song lyrics.

We compare the performance of currently-available versions of ChatGPT (3.5 and 4) given different factors: the method (whether or not the model is forced to pick from a set list of topic labels), the prompt category (whether or not the prompt includes lyrics) and the year of release of each song (pre-2021 or post-2021), which determines whether or not the song could have been included in ChatGPT's training data.

The value of this research stems from the fact that - to the best of our knowledge - no previous scientific study has been carried out on ChatGPT's ability to systematically process and extract information from song lyrics.

Our approach consisted of two steps: first, the selection and creation of our own labelled dataset by using the Songfacts database; second, the implementation of the experiments; and third the comparison of results between ChatGPT3.5 and ChatGPT4. According to our experiments, using ChatGPT3.5 and including the lyrics in the prompt yielded the best results. We also implemented different evaluation methods for ChatGPT given different circumstances, providing initial stepping stones into developing ways to extract information from song lyrics in a (semi-)automatic, scalable way.

Declaration of Authorship

I, Alyssa MacGregor-Hastie, declare that this thesis, titled *Chats, Agents and Lyrics* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a degree degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Date: March 26, 2024

Signed:



Acknowledgments

First and foremost I would like to thank Zoltán Szlávik and Ioannis Petros Samiotis for their unwavering support and guidance through every step of this project. I also want to thank XITE's Data Science team (past and present members) for their companionship and conversations which have been a great inspiration during my internship.

Additionally, I would like to thank Isa Maks and the rest of the faculty at Vrije Universiteit for sharing their knowledge about the complex but fascinating world of Text Mining. Undertaking this Master's course was a big step in a totally different direction, but I am extremely thankful that I did it.

Non potevo non menzionare Giorgio, con cui ho passato ore a studiare, spettegolare e sclerare tra momenti *très bien* e momenti di *kissifregah*- grazie per avermi accompagnato in questa odissea a dir poco avventurosa.

Last but definitely not least, I wouldn't be able to be where I am without the support of Mum, Dad, Liam and Robert, who are unconditionally there for me even when they are tired of my jokes.

List of Figures

3.1	Distribution of genres per Songfacts label	14
4.1	Example format of PGDA. Source: Ding et al. (2022)	21
4.2	Cosine similarity formula	23
5.1	GPT3.5 Confusion Matrix: Knowledge	32
5.2	GPT3.5 Confusion Matrix: Lyrics	32
5.3	GPT4 Confusion Matrix: Knowledge	32
5.4	GPT4 Confusion Matrix: Lyrics	33

List of Tables

1.1	Experiments for prompt-guided topic extraction	4
3.1	10 topic categories selected from Songfacts	12
3.2	Overview of dataset	13
3.3	Quantitative overview of songs	16
3.4	Examples of 5 most frequent words which are identical or similar to the title or topic label of a song	16
3.5	Examples of 5 most frequent words which are different to the title or topic label of a song	17
5.1	Overview of main results for ChatGPT 3.5 and ChatGPT 4	25
5.2	Overlap of correctly classified songs for different prompt category/ChatGPT version combinations	26
5.3	Almost Match songs for Lyrics (Open) - GPT3.5	28
5.4	Almost Match songs for Lyrics (Open) - GPT4	28
5.5	Perfect Match songs for Knowledge (Open) - GPT3.5	28
5.6	Perfect Match songs for Knowledge (Open) - GPT4	29
5.7	Distribution of pre- and post-2021 songs - ChatGPT3.5 and ChatGPT4	29
5.8	Results per class for ChatGPT3.5 and 4	31
5.9	Heartache songs mislabelled as ‘Love’ - ChatGPT3.5 (Knowledge)	33
5.10	Depression songs labelled as ‘Heartache’ - ChatGPT3.5 (Lyrics)	34
5.11	Drugs songs which were mislabelled as ‘Love’ - ChatGPT4 (Knowledge)	35
5.12	Loneliness songs which were mislabelled as ‘Heartache’ - ChatGPT4 (Lyrics)	35
5.13	Depression songs which were mislabelled as ‘Heartache’ - ChatGPT4 (Lyrics)	36
5.14	Cheating songs which were mislabelled as ‘Heartache’ - ChatGPT4 (Lyrics)	36
A.1	Overlapping songs between Knowledge and Lyrics prompts with a Perfect Match score - ChatGPT 3.5 (Closed Method)	46
A.2	Overlapping songs between Knowledge and Lyrics prompts with a Perfect Match score - ChatGPT 4 (Closed Method)	47
A.3	Overlapping songs between Knowledge and Lyrics prompts with a Perfect Match score - ChatGPT3.5 (Semi-Closed Method)	48
A.4	Overlapping songs between Knowledge and Lyrics prompts with a Perfect Match score - ChatGPT4 (Semi-Closed Method)	49

Contents

Abstract	i
Declaration of Authorship	iii
Acknowledgments	v
List of Figures	vii
List of Tables	ix
1 Introduction	1
1.1 Motivation	1
1.2 Research questions	3
1.3 Chapter outline	4
2 Related Work	5
2.1 ChatGPT	5
2.1.1 Transformer architecture	5
2.1.2 Comparing ChatGPT3.5 to ChatGPT4	5
2.1.3 Prompts	7
2.2 Identifying topics in the music domain	8
2.3 Chapter summary	9
3 Ground Truth Data	11
3.1 Data Collection	11
3.2 Overview of dataset	13
3.2.1 Acts	13
3.2.2 Genres	14
3.3 Description of Song Lyrics	15
3.3.1 Structure	15
3.3.2 Language	15
3.3.3 Quantitative overview	15
3.3.4 Most frequent words per song	16
4 Experimental Methodology	19
4.1 ChatGPT Architecture	19
4.1.1 Self-Attention Mechanisms	19
4.1.2 Parameters	20
4.2 Prompts	20
4.2.1 Experimenting with prompts	21
4.2.2 Final selection of prompts	22
4.3 Evaluation	23

4.3.1	Precision, Recall and F1	23
4.3.2	Calculating semantic similarity using cosine	23
4.3.3	Similarity thresholds	24
5	Results & Analysis	25
5.1	Overview of main results	25
5.1.1	Overlapping songs (Closed Method)	26
5.1.2	Closed Method	26
5.1.3	Semi-Closed Method	27
5.1.4	Open Method	28
5.2	Results of Pre- and Post-2021 songs	29
5.3	Results per topic label	30
5.3.1	Results per class	31
5.3.2	Confusion Matrices for the Closed Method	31
5.3.3	Analysis of the Confusion Matrices for the Closed Method	33
5.3.4	ChatGPT3.5 - Knowledge	33
5.3.5	Chat GPT3.5 - Lyrics	34
5.3.6	ChatGPT4 - Knowledge	34
5.3.7	ChatGPT4 - Lyrics	35
6	Conclusion & Discussion	39
6.1	Discussion	39
6.2	Conclusion	41
6.3	Future work	42
A	Appendix	45

Chapter 1

Introduction

Since before the Digital Age, metadata has played a crucial role in the management and expansion of databases and information systems. Metadata is often defined as “data about data”, meaning that it captures various types of information about data such as its context, origin, description, version. Currently, metadata is especially useful to find, access and reuse data items in digital databases, especially as data is generated and collected at a much faster pace.

The same can be said for data in the music domain. As a result of digitisation, music content in the form of songs and videos are becoming increasingly more available through streaming platforms and channels. These platforms rely on metadata to build products (such as featured playlists) and implement recommender systems in order to create a personalised and unique experience for their users. (Chen et al. (2019)).

Typically, in the context of music, metadata can be used to describe audio-related qualities of a song such as mood, tempo and key. However, there is another feature of the song which has been historically overlooked in its potential to contribute to metadata: the lyrics. This is surprising as lyrics present several advantages over other song properties: they are easily accessible through online resources, are non-subjective (i.e. only one version of the lyrics exists) and they convey more meaningful information about the lyrics of the song (such as its topic or its sentiment) (Logan et al. (2004)).

In order to analyse lyrics of songs, some level of expertise and commitment from the annotator responsible would be required - for example, by dedicating time to reading through the lyrics and manually assigning the metadata labels, and by possessing the interpretative skills to assign the correct label to a song. Additionally, music databases are being regularly updated or created with new content which needs new metadata labels. Therefore, a solution needs to be implemented to automatically obtain metadata labels while maintaining the quality of human interpretation.

1.1 Motivation

The motivation behind this project was initiated by XITE Networks, an interactive music video platform which releases its products through ‘channels’ (curated music video playlists), interactive TV apps and on-demand streaming services. XITE’s database includes hundreds of thousands of music videos - each music video is assigned metadata tags relating to the audio features of the song such as mood, key, tempo and genre. The metadata is either received by record labels which own the music video(s) or assigned manually by XITE’s music team: a group of experts responsible for the curation and programming of music videos broadcasted by XITE.

While the input of the Music team is invaluable, relying on them to assign metadata labels is not necessarily the most scalable option, especially while the number of music videos in

XITE’s already extensive database continues to increase.

A solution for this would be to develop an automated method to extract information from song lyrics while maintaining the quality of human interpretation.

Additionally, the metadata present in XITE’s database is tied to the audio content of a song: such as mood, key, tempo and genre. However, the potential of the lyrics’ contribution to metadata generation has been overlooked. As a result, information that is relevant to the song is missing in the database, such as its topic, mood/sentiment and whether or not it contains expletives. This project focuses on extracting the topic of a song based on its lyrics.

Our aim is to provide recommendations to XITE’s Music team for extracting information of a song based on its lyrics. The solution must fulfil the following criteria: it must be scalable, robust and easy-to-use for the Music team. With this in mind and considering the recent developments of state-of-the-art Natural Language Processing (NLP) systems, we decided to use ChatGPT.

ChatGPT is an NLP model created by OpenAI. Since its release in November 2022, ChatGPT has made a significant impact outside and within the Artificial Intelligence community as a result of its ability to generate text in a conversational, human-like manner when prompted by users. As a Large Language Model (LLM), it presents advantages over traditional models: it does not require pre-training or fine-tuning for specific tasks, it is task-agnostic and can handle data that has not been pre-processed. Additionally, it is available as a chat interface, which means it is extremely accessible for users without a technical background.

The release of ChatGPT has triggered a considerable amount of research testing its abilities across different tasks and domains. However, to the best of our knowledge, there is currently no study focused on its ability or behaviour when tasked with identifying and extracting topics from song lyrics.

In NLP, there are three established approaches for extracting topics from text:

- **Supervised:** the model is trained on labelled data and applies the labels that it has learned during training to unseen data (also known as topic classification)
- **Semi-supervised:** the model is trained on data which is partially labelled; the labels in the training data are used to guide the model when it is presented with unseen data (also known as semi-supervised topic modelling)
- **Unsupervised:** the model is trained on unlabelled data and discovers underlying topics from unseen data (also known as topic modelling)

Our project will center around three novel approaches which are inspired by topic classification, semi-supervised topic modelling and topic modelling. The architecture of ChatGPT is different to traditional models in that it does not require training data, generates its output based on prompts and is a black-box system - in other words, details regarding its inner workings and how it processes data are unknown.

As a result of this, we propose to name our task **prompt-guided topic extraction**. Our three proposed approaches for executing prompt-guided topic extraction are as follows:

- **Closed:** whereby ChatGPT assigns labels to instances given a set of labels to choose from (inspired by topic classification)
- **Semi-closed:** whereby ChatGPT is given a choice between a set of labels and freedom to assign its own labels to instances (inspired by semi-supervised topic modelling)
- **Open:** whereby ChatGPT is given complete freedom to assign its own labels to instances (inspired by topic modelling)

We will be repeating the experiment twice by using two types of prompts: **Knowledge** and **Lyrics**:

- **Knowledge:** whereby the model is assumed to know the song in question, therefore only the song title and artist are included in the prompt.
- **Lyrics:** whereby the song title, artist and the lyrics of the song are included in the prompt

The reason that we are using these two prompt contexts is because ChatGPT has been trained on large amounts of textual data obtained from the internet (e.g. Wikipedia articles, books and websites) up to September 2021. Because XITE’s database is regularly updated with new entries (i.e. songs which are released after 2021), we want to test ChatGPT’s ability to process lyrics of songs which it won’t have been exposed to during pre-training.

1.2 Research questions

Having established the above-mentioned focus points, our research will be guided by the following questions and sub-questions:

- **1. Can ChatGPT be used for topic extraction on songs, based on a set of predefined topics vs. a free choice of topics?**
- 1a) To what extent does the release date (pre- or post-2021) of the song affect ChatGPT’s ability to do topic extraction on songs?
- 1b) How does including the lyrics in the prompt affect ChatGPT’s ability to do topic extraction on songs?
- 1c) To what extent does the version of ChatGPT (3.5 or 4) affect its ability to perform prompt-guided topic extraction?

In order to answer these questions, we will adopt the following steps: first, we will create our own dataset with 100 songs collected from the Songfacts website (more of which will be outlined in the Data section); the final dataset will comprise 10 songs for 10 topic labels. These 10 songs will be split into two categories: 5 songs will be from pre-2021, and 5 songs will be from post-2021. Therefore, the dataset will have 50 songs which are pre-2021 and 50 songs which are post-2021.

Following this, we will run two versions of ChatGPT (3.5 and 4). For both versions, we will be carrying out a total of six experiments, which are represented in the following table:

Once we have obtained the results, we will evaluate the output of both ChatGPT versions by applying cosine similarity between the ground truth label acquired by Songfacts and the predicted labels from each approach and prompt category.

The similarity scores will give an indication of how ChatGPT’s behaviour is affected by the given prompt and approach. For example, we might observe that, if the model is presented with lyrics and is forced to choose a label from a given set of labels (Closed method), then the predicted label(s) will obtain a higher similarity score compared to another method and/or prompt.

Additionally, we will observe the distinction (if any) between the results from songs dating pre- and post-2021. For example, we might observe that pre-2021 songs score overall higher compared to post-2021, given that ChatGPT will have access to more information related to the songs and the associated artist(s) from pre-training.

In tandem with this, we will be observing and comparing both versions (3.5 and 4) in their ability to carry out prompt-guided topic extraction. While ChatGPT4 presents some

Number	Prompt	Data	Approach
1	Knowledge	50 Pre-2021 songs	Closed
		50 Post-2021 songs	
2	Knowledge	50 Pre-2021 songs	Semi-Closed
		50 Post-2021 songs	
3	Knowledge	50 Pre-2021 songs	Open
		50 Post-2021 songs	
4	Lyrics	50 Pre-2021 songs	Closed
		50 Post-2021 songs	
5	Lyrics	50 Pre-2021 songs	Semi-Closed
		50 Post-2021 songs	
6	Lyrics	50 Pre-2021 songs	Open
		50 Post-2021 songs	

Table 1.1: Experiments for prompt-guided topic extraction

advantages over ChatGPT3.5 (e.g. it has been trained on a considerably larger amount of data), it requires a paid subscription and has a usage cap; this can be a potentially limiting and costly solution. Due to these limitations, we are motivated to find out if using the freely accessible version (3.5) would achieve comparable results to the newer version.

1.3 Chapter outline

In this chapter we provided the context and motivation for our project. The rest of this thesis will be structured as follows: **Chapter 2** focuses on Related Work; **Chapter 3** outlines the collection process and description of the ground truth (Songfacts) data; **Chapter 4** provides the methods for the experiments; **Chapter 5** presents the results and analysis of our experiments; **Chapter 6** concludes with discussion points and recommendations for future implementation.

Chapter 2

Related Work

To the best of our knowledge, there is currently no research surrounding ChatGPT’s ability to perform topic extraction on song lyrics. However, we can refer to certain resources which cover ChatGPT and different methods of extracting topics in the music domain.

2.1 ChatGPT

In this section, we will touch on transformer architecture, which ChatGPT is built on. Following this, we will explore studies which follow the evolution from GPT3 to ChatGPT4. This leads to a comparison between ChatGPT3.5 and ChatGPT4, along with an overview of papers which also uncover usages of prompt engineering.

2.1.1 Transformer architecture

GPT is built on transformer architecture, which was designed to address the limitations of traditional sequence-to-sequence models, like recurrent neural networks (RNNs) and their variants, which struggle with capturing long-range dependencies in sequences. Transformers use a self-attention mechanism to capture these dependencies effectively and process sequences in parallel, making them highly efficient for NLP tasks. The transformer architecture consists of two main components: the encoder and the decoder. Both GPT3.5 and GPT4 are decoder-only transformers, and can therefore be used for autoregressive language generation tasks.

Although GPT3.5 and GPT4 are built on the same architecture, they are said to differ considerably in terms of their size. OpenAI released an official technical report to coincide with the release of GPT4, however there is little information pertaining to important elements such as the parameters and the data on which the model has been trained. In fact, the report omits any mention of the parameters and makes a passing mention of the training data as “publicly available data (such as internet data) and data licensed from third-party providers.” (OpenAI (2023a)) As a result, it is difficult to make a quantifiable comparison of GPT3.5 and GPT4. Regardless of this, it has been universally acknowledged that GPT4 is far bigger in size compared to its predecessor, especially given its ability to process multimodal material and understand multilingual input.

2.1.2 Comparing ChatGPT3.5 to ChatGPT4

In order to gain an understanding behind the conception of ChatGPT, we first refer to research evaluating the capabilities of earlier versions: GPT3 and GPT3.5. Brown et al. (2020) state the limitations of certain so-called task-agnostic models which still require a task-specific dataset and further fine-tuning in order to perform a specific NLP task. Additionally,

they state their aspiration to create a system which obtains the same ‘fluidity and generality’ of humans, which are able to handle most language tasks when prompted with a short question or demonstration. As a result of this, GPT3 is introduced and evaluated over a number of different NLP tasks such as Question-Answering, Translation and Reading Comprehension without any additional pre-training or fine-tuning.

While GPT3 achieved a strong performance across these tasks, Brown et al. (2020) observe some limitations as a result of their experiments. One notable mention was that of GPT’s architecture as an autoregressive model, which posed a challenge to tasks which would benefit from bidirectionality. This includes tasks that would require filling-in-the-blanks and tasks that require generating a short answer after analyzing a long passage. The latter task is especially relevant to our research, as we will be evaluating the model’s ability to provide a one-word description of a song based on its lyrics. The paper also expresses uncertainty about whether or not the model learns new tasks during inference, or if it recognizes and identifies tasks that it has learned during training.

By contrast, Liu et al. (2021) go as far as calling GPT3 a ‘miracle’ and claim that it can perform as well as a bidirectional model such as BERT in terms of natural language understanding. However, one significant limitation concerning a giant model like GPT3 is its poor transferability as a result of its size, meaning that it is not able to memorise fine-tuning samples that would benefit a particular task. This claim is supported by Modari et al. (2021), who test GPT3’s capabilities to perform NLP tasks using biomedical text corpora. From their research, GPT3 appears to significantly underperform compared to other tasks outside the biomedical domain.

Following these researches and in the run-up to GPT4, GPT3.5 was released as a first step to address the limitations of GPT3. Rather than releasing GPT3.5 in its entirety, this version became available in a number of different versions created for specific purposes. Zhao and Zhou (2023) present a comparative analysis of GPT3 and GPT3.5 (amongst others) in the paper as stated in ‘A Survey of Large Language Models’. According to their findings, what distinguishes GPT3.5 from GPT3 is the fact that GPT3.5 is trained on code as well as text in order to achieve improved performance in reasoning tasks involving code and arithmetic problems. Although these tasks are not specifically tied to understanding natural language, GPT3.5 is reported to perform NLP tasks at a higher level compared to its predecessor.

Despite these accolades, certain research alludes to being less convinced about GPT3.5’s abilities. Ye and Chen (2023) compare the performance various versions of both models on tasks such as Machine Translation, QQP, Named Entity Recognition and Sentiment Analysis. Surprisingly, the newest interface tested - gpt-3.5-turbo - only achieved the highest score for a limited number of the tasks. Additionally, a question is raised about the robustness of the gpt-3.5-turbo model which appears to have improved very little (if at all) compared to older versions of GPT. With this, Ye and Chen (2023) reach the conclusion that newer and bigger models are not necessarily the best option for all tasks. In fact, they go as far as to claim that GPT3.5’s increased ability to generate human-like responses may compromise its ability to successfully perform other tasks. This statement is of particular interest to us, as we will be assessing GPT3.5’s performance on a task which requires very minimal conversation - if anything, we are expecting it to act as a model rather than a human. We will now gain some preliminary understanding of the abilities and limitations of GPT4 (a newer and bigger model) by exploring literature which supports and/or contradicts claims made in this section.

A few months after the release of ChatGPT (which was based on GPT3.5), GPT4 was made available through ChatGPT Plus, a premium version of ChatGPT which is only accessible through paid subscription. Just as GPT3.5 was updated by being able to process code, GPT4 is also able to process multi-modal input (e.g. images)

As a result of the initial buzz surrounding ChatGPT/GPT3.5, copious amount of research

has already been (and continues to be) published in regards to the abilities (or lack thereof) of GPT4. Because a waiting list was implemented in order to obtain the API for GPT4, a majority of research has been made using GPT4 through its chat version. This has resulted in additional insights concerning the role of the prompt in producing the desired output of the model.

2.1.3 Prompts

Liu et al. (2023)'s study uses ChatGPT/GPT4 for the task of de-identification in the medical domain, where prompts are utilised to delete confidential information while preserving the meaning and structure of clinical reports. Additionally, they compare the ChatGPT/GPT4 to the ChatGPT/GPT3.5, as well as BERT, RoBERTa and ClinicalBERT. Out of all the models, GPT4 achieves the highest accuracy rate (over 0.99) - an impressive result, especially when compared to a fine-tuned model in the medical domain such as ClinicalBERT. Along with these insights, the study carries out an extensive analysis of the model by assessing its behaviour when presented with different types of prompts - through these observations, certain limitations of the model are highlighted. The authors identify 'Bad Prompts' in which GPT4 struggles to execute the task correctly as a result of the presence of additional/unnecessary punctuation, failure to specify the desired output or stating multiple tasks in one instance. As a result, the concept of prompt engineering is brought up as an important technique to optimize results and regulate the behaviour of LLMs.

A study carried out by Nori et al. (2023) enforces claims of GPT4's superior performance over other models which are specifically fine-tuned for data in the medical domain, as well as GPT3.5. However, these results were achieved 'without any specialized prompt crafting' - meaning that very little emphasis is given on the type of prompt used to obtain their results. Despite this, the authors take advantage of GPT's conversational abilities by asking it to provide its reasoning and explanations behind its answers. This proves to be a very useful step in the evaluation process, as it not only can provide a qualitative insight into its output, but can also shed light onto limitations such as inaccurate recommendations regarding diagnoses and testing, hallucinations and factual errors.

Another interesting observation is presented by Ali et al. (2023), where GPT4 and GPT3.5 are evaluated on their ability to provide correct answers to a 500-question mock neurosurgical examination. While both models received a passing grade, GPT4 was found to outperform GPT3.5 across all twelve question categories. Although these findings are not surprising at this point, the authors noted that GPT3.5's accuracy particularly floundered when presented with questions of greater word length compared to GPT4. This observation is quite relevant to our research, as the length of the prompts belonging to the Lyrics category will be considerably longer compared to prompts that GPT is usually used to process. We could therefore make a preliminary hypothesis that GPT4 could perform better in the Lyrics category compared to GPT3.5.

The research discussed in this section highlights the diverse approaches taken when implementing and evaluating the performance of LLMs. Therefore, no uniform or 'one-size-fits-all' method has emerged for successfully executing a given task.

Some researchers emphasize the significance of prompt engineering to achieve desired results. They invest considerable effort in crafting well-structured and targeted prompts to elicit specific responses from the LLM. By carefully curating the input, they can fine-tune the model's outputs and optimize its performance for particular tasks or domains.

On the other hand, some studies adopt a more minimalistic approach to prompts, choosing to invest less effort in prompt engineering. Their aim is to assess the LLM's natural language processing capabilities without heavily guiding or biasing its responses. In doing so, they seek a broader understanding of the model's general competence across various domains.

A common observation across these studies is the trade-off that arises as models or model versions increase in size. As GPT increases in size, it tends to exhibit improved performance in certain tasks due to the increased capacity and learning capabilities. However, this enhancement can come at the cost of performance on other tasks, where the model might show little or no improvement or even degrade in accuracy.

As we can deduce from this section, a considerable amount of research has been covered concerning ChatGPT4's capabilities in the medical domain. In the next section we will review some studies that are relevant to our domain of interest: music.

2.2 Identifying topics in the music domain

As previously stated, there is currently no research surrounding GPT's ability to perform topic extraction on song lyrics. However, previous studies have already been carried out on traditional models and their ability to process song lyrics through supervised and unsupervised methods (topic classification and topic modelling, respectively). In this section, we will explore how these two methods have been used in the context of the music domain.

First, we will take a deeper look into unsupervised methods. In relation to our research, this approach is most similar to the Open Method, where GPT will be given the choice to assign all data instances with a label of its own choice - therefore, we will have no expectation or idea of what labels it will generate and will be difficult to evaluate on a quantitative level.

Two popular topic modelling methods are Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF). Research has already been done on various topic models' abilities to process song lyrics, such as Latent Dirichlet Allocation(LDA) (Denzler (2021)) and Non-negative Matrix Factorization (NMF) (Kleedorfer et al. (2008)). As there is no standardised way of evaluating the performance of unsupervised methods, both studies carry out different approaches to fit their purpose. Denzler (2021) makes use of gensim's *Cv* module in order to measure the topic coherence for each cluster outputted by the LDA model - this is because this particular module obtained the highest correlation with human ratings and was therefore considered to be the most reliable. The LDA's performance is therefore assess the coherency of the topic clusters. Kleedorfer et al. (2008) adopt an approach similar to inter-annotator agreement in order to facilitate the evaluation process for the output of the NMF model, which takes place in two phases. In the first phase, subjects are presented with the most relevant terms of each outputted cluster and asked to provide labels to summarise/describe each group of terms. In the second phase, the same term groups are shown to the same test subjects, who are then asked to choose the most suitable labels that were generated during the first phase. A probability calculation is then implemented to determine the agreement between subjects and therefore the coherence of a topic cluster generated by the NMF model. As we can observe from these studies, an unsupervised learning method might entail a higher level of human intervention compared to other methods (namely supervised learning); meaningful insights might therefore only be obtained if a qualitative approach is implemented. As a result, we might want to consider involving XITE's Music team to observe the extent to which the predicted labels from the Closed method align with human judgement, and therefore gain better understanding of GPT's performance given a particular prompt context.

Conversely to topic modelling, topic classification makes use of labelled data in training; as a result, evaluation is more straightforward as one can make use of standard metrics such as Precision, Recall and F1 to determine how successful the performance of a model is. In relation to our research, topic classification is most similar to our Closed Method, whereby GPT will be forced to choose a label from a set list to assign to instances.

For this section, we will be taking a closer look at studies carried out by Papazoglou and Gaizauskas (2021) and Choi et al. (2014), who both explore the effect of song lyrics

on models' abilities to perform topic classification. These two studies are also of particular interest to us because they both make use of Songfacts.com, from which we will also be extracting songs and their respective lyrics to create our dataset. It is important to note that both of these studies also take into consideration the user-generated interpretations for each song - which are also available on Songfacts - as a feature on which to train the classifiers. This is relevant to our research because our Knowledge prompt will rely on GPT's knowledge of a song based on the internet data that it has been trained on, and therefore could already have been exposed to the interpretation of the song in question. This could facilitate eventual comparisons between the performance of GPT and traditional models for this specific task and domain for both the Knowledge and Lyrics contexts.

As a first step in their research, Papazoglou and Gaizauskas (2021) create their own dataset by extracting 130 songs and their respective lyrics from the 20 most populated topic categories from Songfacts, resulting in a balanced set of 2,600 songs. Following this, four classifiers - Logistic Regression, Multinomial Naive Bayes, Random Forest and k-NN - are run on this dataset with five different feature combinations, including using Lyrics and Interpretations as standalone features. Interestingly, these two single features prove to be the two lowest-scoring feature categories, with the Interpretations obtaining a minimal advantage over Lyrics.

In order to gain understanding into the low scores obtained from these features, a more thorough analysis is carried out by examining potential confusions between topic categories. Some of the frequently misclassified categories were identified as Heartache, Breakup, Ex-partner and Cheating. As we will be using the categories Heartache and Cheating in our dataset, we are interested to find out if similar instances of confusion will also occur.

Choi et al. (2014) carry out a similar study by evaluating and comparing the role of lyrics and user-generated interpretations on topic classification, using songs collected by Songfacts and running their model on a balanced dataset (900 songs, 90 songs per 10 topic categories). The output of their experiment (carried out on a K-NN model) supports Papazoglou and Gaizauskas (2021)'s claim that using only the Lyrics as a feature obtains the lowest-scoring performance for the model. Following this, further insights are obtained from identifying the most frequently misclassified classes: Old Girl/Boyfriend, Loneliness, Heartache, and Cheating. The authors single out the classes Loneliness and Heartache as being frequently confused and argue that their shared negative mood/sentiment could be a reason for their misclassification.

As Choi et al. (2014) and Papazoglou and Gaizauskas (2021) have obtained corresponding insights and use some of the same labels that we will be featuring in our dataset (Loneliness, Heartache and Cheating), we are curious to find out how the behaviour and output of an LLM such as GPT compares to that of the traditional models exhibited in these two studies. While assigning metadata labels based on the mood/sentiment of a song is not our focus for this particular research, it will be interesting to observe how (and if) an LLM is able to distinguish between topic categories which share similar emotional characteristics.

2.3 Chapter summary

In this chapter, we presented studies which enabled us to make comparisons between ChatGPT3.5 and ChatGPT4, along with gaining insights into how a role affects the performance of the model. Following this, we explored how traditional methods were used to identify topics in the music domain. We particularly focused on two studies which, like us, use Songfacts to create their own labelled dataset. This also allowed us to make some preliminary hypotheses into the performance of ChatGPT on song lyrics.

Chapter 3

Ground Truth Data

This chapter covers the data collection process, an overview of the acts and genres in the dataset and a description of the song lyrics.

3.1 Data Collection

To create our dataset, we collected data from Songfacts (<https://www.songfacts.com/>). Songfacts is a searchable database which features various types of information about songs such as:

- Song Lyrics: The complete lyrics of the song.
- Songfacts: These are the interesting or lesser-known facts and stories behind the song, including details about the song’s creation, inspiration, and any hidden meanings or anecdotes related to it.
- Songwriter and Artist Information: Details about the songwriters, composers, and performers of the song.
- Song Genre and Style: Information about the musical style and genre of the song.
- Release Date and Chart Performance: Information about when the song was released and how well it performed on music charts.
- Album Information: Details about the album on which the song appears.

Most importantly, Songfacts has an ‘About’ section with 219 entries - each entry is a topic category (<https://www.songfacts.com/category/type-about>). Example of these topic categories are: ‘Songs about being free’, ‘Songs about childhood’, ‘Songs about dogs’. We collected all the songs for our dataset from this section.

The steps taken to collect the data are as follows:

- **1. Create the data document.** We started by creating a CSV file with columns: Track Title, Artist, Lyrics, Pre-2021, Songfacts Label (or ground truth label).
- **2. Define the number of songs and topic categories.** As training data was not needed and we wanted to maintain the quality of the data as high as possible, we decided to use a maximum of 10 songs per 10 topics, summing up to 100 songs in total.
- **3. Define the extent of popularity of the artist/songs to be collected.** In order to observe GPT’s behavior when presented with pre-2021/post-2021 music, we needed to define the extent of popularity of the artists and songs to be included. Our aim

was to ensure that GPT had enough knowledge of the acts to make fair and educated predictions about their songs. To achieve this, we decided to focus on songs associated with popular acts. To aid our search for popular acts, we consulted reputable sources like the Rolling Stone article ‘200 Greatest Singers of All Time’ and the Billboard article ‘Greatest All Time Artists’.

4. Select the topic categories We referred to Papazoglou and Gaizauskas (2021)’s report which listed the 20 most populated topic categories from Songfacts. It can be assumed that Papazoglou and Gaizauskas (2021) selected the topic categories with the most number of songs in order to obtain enough data on which to train their models - in our case, we only picked topic categories that had enough songs that were released before and after 2021. The final list of selected topic categories can be found in the following table:

Love	Heartache
Friendship	Cheating
Depression	Death
Drugs	Loneliness
Sex	Religion

Table 3.1: 10 topic categories selected from Songfacts

- **Select the songs/acts** Once we selected the topic categories, we manually inspected the song list for each topic and selected the songs that fit our criteria. In some cases, there were songs that had more than one topic label; although we did not let this affect our selection process, we will take this into consideration if we see any anomalies in our results.

To ensure that the selected artists were popular, we cross-referenced the artists in the Songfacts topic categories with the Rolling Stone and Billboard article. However, during this process we noticed that some prominent acts, such as Red Hot Chili Peppers, Metallica, and Green Day, were missing from the charts.

Despite these omissions, we chose to include these acts in our dataset. To validate the popularity of the listed artists and ensure a high-quality selection of songs, a member of the Music team reviewed the final song selection.

- **5. Add the relevant song information and its lyrics to the CSV file.** There was no preprocessing step, therefore features such as stop words, punctuation and expletives are all included in the lyrics. Below is an example and description of an entry in the dataset:

Track	Artist	Lyrics	Pre-2021	Songfacts Label
All I Can Do	Dolly Parton	Well, it’s all I can do	Yes	Love

- **Track:** The title of the song
- **Artist:** The artist associated to the song
- **Lyrics:** The text of the song
- **Pre-2021:** ‘Yes’ or ‘No’ is used to indicate whether or not the song was released before 2021
- **Songfacts Label:** The topic label assigned by Songfacts

3.2 Overview of dataset

In this section, we will present an overview of the acts and genres present in our dataset.

3.2.1 Acts

Our dataset comprises a combination of solo artists, collaborations and bands. For this reason, we are categorising all these entities under the term ‘act’ to avoid potential confusion.

Total number of acts	66
N. of most frequent acts	5, 4 and 3
N. of bands	18
N. of collaborations	4
N. of individual acts	44
N. of male acts	41
N. of female acts	24
N. of mixed acts	1
N. of official popular acts	60
N. of unofficial popular acts	6
N. of topic categories with 10 unique acts	6
N. of topic categories with duplicate acts	4

Table 3.2: Overview of dataset

- **Total number of acts:** As some acts are repeated throughout the dataset, we only counted the total number of unique acts, which is 66 out of 100.
- **Number of most frequent acts:** The number of acts which featured the most times in the dataset. These are: Ed Sheeran (5 times), Taylor Swift (5 times), Beyoncé (4 times) and Adele (3 times).
- **Number of bands:** The total count of bands in the dataset. In our case, a band is defined as a musical group which regularly releases music under the same name. For example: Metallica and Green Day.
- **Number of collaborations:** The total count of collaborations in the dataset. In our case, a collaboration comprises two artists which release a song as a one-off occasion. For example: Elton John ft. Charlie Puth.
- **Number of individual acts:** The total count of individual acts in our dataset. By this, we mean either a single artist which regularly releases music under the same name (e.g. Adele) or a member of an existing band which has released music as an individual act (e.g. Chris Cornell from Audioslave or Kurt Cobain from Nirvana).
- **Number of male acts:** The total count of acts (band or individual) which identify as male. For example: Green Day or Frank Sinatra.
- **Number of female acts:** The total count of acts (band or individual) which identify as female. For example: Katy Perry or Destiny’s Child.
- **Number of mixed acts:** The total count of band or collaborations which features members of opposite genders. In our dataset, we only had one: the collaboration between Billie Eilish (female) and Labrinth (male).

- **Number of official popular acts:** The total count of acts which have been featured in the Rolling Stone article: ‘200 Greatest Singers of All Time’ and the Billboard article: ‘Greatest All Time Artists’. We will expand on this in the next section.
- **Number of unofficial popular acts:** The total count of acts which are not featured in the abovementioned articles, but can still be regarded as popular.
- **Number of topic categories with 10 unique acts:** Out of 10 topic categories, there are 6 which feature 10 songs by 10 different acts: Love, Friendship, Cheating, Religion, Death and Loneliness.
- **Number of topic categories with duplicated acts:** There are 4 topic categories which feature an act which is included at least twice. These are: Sex (Beyoncé - 3 times), Depression (Ed Sheeran - 2 times), Heartache (Taylor Swift - 2 times), Drugs (Eminem - 2 times).

3.2.2 Genres

In order to obtain the genre for each song, we referred to the Music Genre Finder tool from Chosic (<https://www.chosic.com/music-genre-finder/>). This tool uses Spotify and Wikipedia as references for providing a list of genres (and sub-genres) for each song in our dataset. Because Wikipedia is an open-source platform, we decided to only use Spotify as it would be more reliable.

From consulting the Spotify lists, we realised that a large majority of the songs in our dataset had numerous genre tags. For example, ‘Tippa My Tongue’ by Red Hot Chili Peppers had the following: Rock, Permanent Wave, Funk Rock, Funk Metal and Alternative Rock. In the interest of simplifying the process, we decided to classify each song under the following ‘main’ genre tags: Country, Jazz, Pop, R&B, Rap, Rock and Soul. We then used the Spotify genre tags to find the main genre that would be most relevant to each song. For example, because ‘Tippa My Tongue’ had 3 out of 5 genre tags which featured the term ‘rock’, we decided to assign it to the Rock genre category.

Below is an overview of the distribution of genres per Songfacts label:

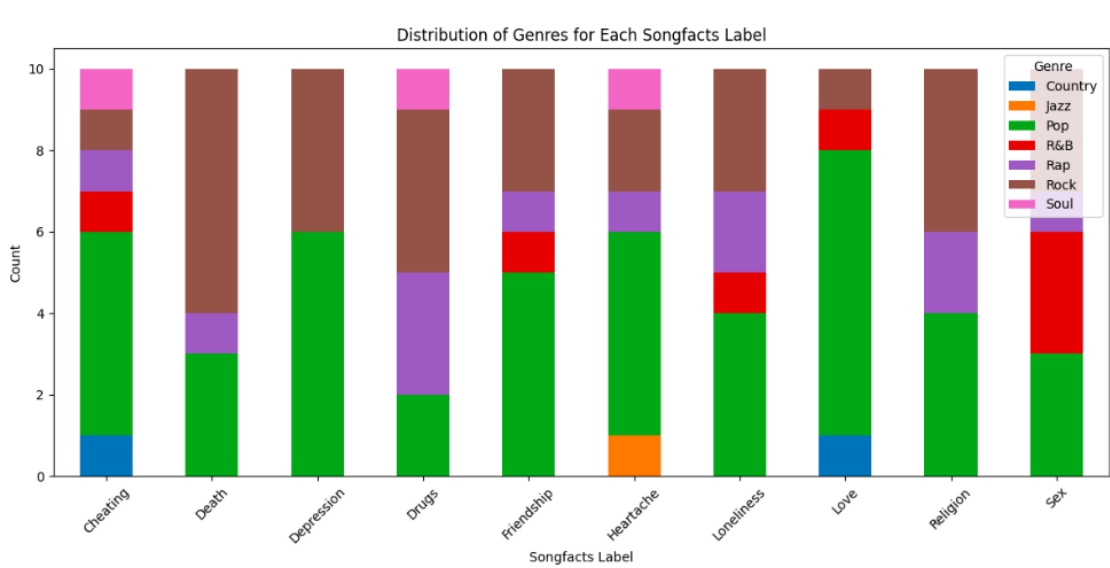


Figure 3.1: Distribution of genres per Songfacts label

From this overview, we can observe that the two most frequent genres are Pop and Rock.

In terms of topic categories, Love has the highest number of Pop songs (7 out of 10), while Death has the highest number of Rock songs (6 out of 10).

On the other hand, Jazz and Country are the two genres that occur the least often. Country is only featured twice across all songs (once for Cheating and once for Love); Jazz is only featured once, for the Heartache topic category.

3.3 Description of Song Lyrics

In this section, we will expand on the structure, language and lyrical content of the songs in our dataset.

3.3.1 Structure

The songs in the dataset are in verse form, which is different from prose form. Unlike prose, which relies on sentences and paragraphs for its organization, verse form employs lines and stanzas that adhere to a specific metric structure, wherein rhyme and rhythm play a significant role. Overall, the songs are in a verse-chorus structure, where some sections of the song are repeated for emphasis.

In our dataset, the songs are laid out in prose form: in other words, there will be no new lines or gaps between each line or stanza. However, the first character of each line is capitalized - therefore we assume that ChatGPT should be able to differentiate between the end and beginning of a line.

3.3.2 Language

All of the songs present in our dataset are originally written in English. During the collection process, we realised that Songfacts features English translations of songs (such as ‘Flower’ by Jisoo in Korean) and some which comprise a combination of different languages (such as songs by Rosalia, which often combine English with Spanish). Following this discovery, we decided to avoid including these songs as they might affect the model for the following reasons:

1. Nuances and idioms that could be present in the original song might be lost or overlooked in the translated version, which could prevent the model from making a sensible interpretation. Additionally, we can assume that the model might already have been exposed to the song in its original language during pre-training. Therefore, when implementing the Knowledge prompt, it is likely that ChatGPT will refer to the original song, rather than to its English translation.
2. Although ChatGPT is able to process multilingual input, there is a chance that it might get confused if presented with a song which features more than one language. This could therefore prevent the model from interpreting the lyrics to its full potential.

3.3.3 Quantitative overview

Below is a quantitative overview and description of the song lyrics in our dataset:

- **Highest number of total words:** ‘Jesus Lord’ by Kanye West has the highest number of total words (including stop words and repeated words) out of all the songs. This means that it is also the longest text excerpt that ChatGPT has to process.
- **Lowest number of total words:** ‘Ghosts Again’ by Depeche Mode has the lowest number of total words (including stop words and repeated words) out of all the songs. This means that it is also the shortest text excerpt that ChatGPT has to process.

Highest n. of total words	740
Lowest n. of total words	43
Average n. of words	116.18
Highest n. of unique words	459
Lowest n. of unique words	26
Average n. of unique words	94.65

Table 3.3: Quantitative overview of songs

- **Average number of words:** The average number of words in a song, including stop words and repeated words.
- **Highest number of unique words:** ‘Jesus Lord’ by Kanye West has the highest number of unique words (i.e. each individual word is only counted once, including stopwords) out of all the songs.
- **Lowest number of unique words:** ‘40’ by U2 has the lowest number of unique words out of all the songs.
- **Average number of unique words:** The average number of unique words in a song i.e. each word is only counted once - this also includes stopwords.

3.3.4 Most frequent words per song

As an additional step, we also extracted the 5 most frequently occurring words in each song. To do this, we used the NLTK library to remove stopwords and implemented *Counter* from the Collections module to identify and count the occurrence of each word. From this process, we observed that some songs had frequently occurring words which were identical or very similar to their title or topic label; others had frequently occurring words which were not related or at all similar to the title or topic label. Below are some examples:

Title	Act	Label	5 most frequent words
Crazy Little Thing Called Love	Queen	Love	‘thing’, ‘called’, ‘yeah’, ‘little’, ‘love’
Never Felt So Alone	Billie Eilish ft Labrinth	Loneliness	‘alone’, ‘felt’, ‘never’, ‘na’, ‘oh’
Act of God	Prince	Religion	‘act’, ‘want’, ‘godcall’, ‘except’, ‘god’
Best Friends	The Weeknd	Friendship	‘yeah’, ‘best’, ‘friend’, ‘oh’, ‘friends’

Table 3.4: Examples of 5 most frequent words which are identical or similar to the title or topic label of a song

Title	Act	Label	5 most frequent words
Bobby Jean	Bruce Springsteen	Friendship	'could', 'wished', 'ever', 'say', 'goodbye'
Eddie	Red Hot Chili Peppers	Drugs	'remember', 'last', 'night', 'i'm', 'please'
Rich Spirit	Kendrick Lamar	Religion	'duh', 'i'm', 'ah', 'dun', 'ooh'
Lithium	Nirvana	Depression	'i'm', 'gonna', 'yeahyeah', 'cracki', 'cause'
Confession	Destiny's Child	Cheating	'ooh', 'clean', 'oh', 'feel', 'uh'

Table 3.5: Examples of 5 most frequent words which are different to the title or topic label of a song

In Table 3.4, we can see how the five most frequently occurring songs are identical or very similar to the title and topic label of the song. For example: for 'Best Friends' by The Weeknd, three of the most frequent words are 'best', 'friends' and 'friend'; it would therefore be relatively easy to identify the correct topic label of the song (Friendship).

In Table 3.5, we can observe the opposite. For example: 'Bobby Jean' by Bruce Springsteen also belongs to the topic label Friendship. However, none of its five most frequently occurring words can be linked to the concept or topic of Friendship.

In the case of both tables, the majority of the songs have at least one frequent word which is specific to the music domain: such as 'oh', 'yeah', 'ah' and 'uh'. By choosing to not remove them from our dataset, we would be able to gain insights into whether or not ChatGPT will be affected by the presence of these domain-specific terms.

Chapter 4

Experimental Methodology

In this section, we outline characteristics of ChatGPT’s architecture, along with our process for creating our prompts and evaluation methods.

4.1 ChatGPT Architecture

ChatGPT is based on transformer model architecture (Vaswani et al. (2017)). The transformer model was designed to overcome some of the limitations of earlier neural network architectures, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs). Perhaps most importantly, the core innovation of the transformer model is the self-attention mechanism. In this section, we will outline how self-attention mechanism will benefit our task, in addition to listing the parameters that we will use.

4.1.1 Self-Attention Mechanisms

ChatGPT is built on self-attention mechanisms, which enables it to excel in three areas: context comprehension, handling long-range dependencies and selectively focusing on different parts of a given input (or in our case, a prompt). As a result, it is accessible to users of a non-technical background and also ideal for understanding and interpreting text in a similar way to humans.

Context comprehension: In the case of the Lyrics prompt, which is considerably long given that it features the lyrics of 100 songs, it is important for the model to identify and capture relationships between key words in order to make an informed prediction. The prompt is relatively unstructured: in our case, we copied and pasted the song title and lyrics directly into the ChatGPT window. As a result, there is no gap between songs and partial distinction between the song title and the lyrics of the song. Despite this, ChatGPT is able to distinguish each song and generate an output as required in the prompt. This makes it very accessible to use for XITE’s Music team as there will be no need to preprocess or restructure the data.

Additionally, because the prompt comprises specific instructions to execute the task, it is important for the model to understand what is required of it. If we want the Music team to eventually use GPT in their metadata-extraction process, we need a model that is able to carry out a task by understanding natural language. Additionally, song lyrics can have hidden meanings or more subtle nuances, therefore it is important to capture relationships between words.

Handling long-range dependencies: Attention mechanisms assist GPT models in managing long-range dependencies in text, where the connotation of a word can be influenced by another word situated elsewhere in the text. This can be useful, as a word in the verse might have a less ambiguous meaning if it is connected to a word in the chorus, for example.

4.1.2 Parameters

As with any model, parameters can be adjusted to fit a specific task. However, because GPT has not yet been used for topic extraction on lyrics, it is not known which parameter settings work best. Therefore, we will carry out our experiments using ChatGPT's default settings.

- **max tokens** - this setting determines the maximum number of tokens that GPT is allowed to generate in its response. As we will be making use of GPT's conversational skills by asking it to provide explanations for some of its predictions, we will be maintaining the default number of 2048.
- **n** - this setting indicates the number of responses required for a given prompt. For example, if n is set to 3, then GPT will generate 3 different responses to a single prompt. This might be an interesting parameter to experiment with - for example, to see if a song is assigned different topic labels per new response. However, given the scope of the project, we will keep the parameter at $n = 1$.
- **stop** -The stop parameter is useful for setting a limit to the length of GPT's response. As we are prompting GPT to assign labels to a determined number of songs, we do not expect it to keep generating words once it has completed its task. For this reason, we will maintain the default setting of *None*.
- **temperature** - this setting determines how inventive and creative the output of GPT can be, given a number between 0-1. The higher the number, the more creative/diverse the response will be. While we want to give GPT a chance to use its interpretative skills, we also want to prevent it from generating a response that is too distant from what is required. For this reason, we maintained the default setting of 0.5

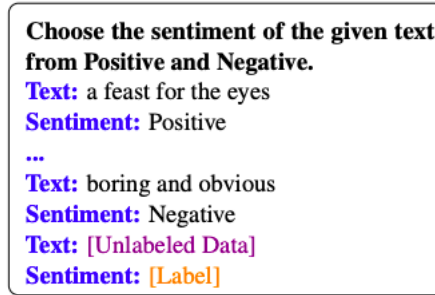
4.2 Prompts

As observed in the Related Work section, there is (currently) no one-size-fits-all approach for designing prompts to fit a specific purpose. At the time of writing, a considerable amount of attention has been placed on the concept of Prompt Engineering i.e. the intentional design and crafting of input prompts that are given to large language models in order to guide and generate specific responses.

While reviewing resources that would guide us to crafting the best prompt, we realised that a large majority of them were focused on the best approaches for generating long or conversational-like responses. As we also mentioned in the Related Work, we are expecting GPT to act more like a model rather than a human by manipulating it to generate a single word - the topic label - rather than an extensive text excerpt. As a result of this, we redirected our search and decided to base our process on Prompt-Guided Unlabelled Data Annotation (PGDA). In this approach, GPT takes on the role of a data annotator through task-specific prompts which force it to assign labels to unlabelled data instances. An example format of PGDA provided by Ding et al. (2022) is as follows:

From this example we defined the three main components that would help us construct our own prompt(s) to best fit our purpose:

1. **Task:** The specific task required from ChatGPT
2. **Example:** A one-shot example of the kind of output we expect from the model
3. **Unlabelled Data:** The data instances that we require ChatGPT to label



Choose the sentiment of the given text from Positive and Negative.
Text: a feast for the eyes
Sentiment: Positive
 ...
Text: boring and obvious
Sentiment: Negative
Text: [Unlabeled Data]
Sentiment: [Label]

Figure 4.1: Example format of PGDA. Source: Ding et al. (2022)

4.2.1 Experimenting with prompts

In order to determine the optimal prompts to carry out topic extraction on song lyrics, we carried out an experimentation phase to help us understand which approach would yield the desired output from the GPT models. This process was carried out using the GPT3.5 version of ChatGPT. This is because we assumed that if the prompts work for GPT3.5, then they would also work for GPT4.

Below we outline some observations made during the process:

- In order to prevent potential confusion and to ensure the topic extraction process be as manageable as possible, we found it best to have a separate ChatGPT window per prompt (Lyrics or Knowledge), Method (Closed, Semi-Closed, Open) and topic category (e.g. ‘Love’ or ‘Friendship’). For example, we would have one window for ‘Knowledge (Closed) Love’, one window for ‘Knowledge (Closed) Friendship’ etc. Additionally, having separate chat windows would prevent the model from having its judgment affected by earlier conversations and would therefore allow for less biased predictions.
- Also in the interest of manageability, it is preferable (and possible) to include all the songs in the prompt at once, rather than singling out one song at a time. We found that for the Knowledge prompts, it was possible to list all 10 songs at the beginning of the prompt without any initial context while still obtaining the desired output. Conversely, for the Lyrics prompts, we placed the song lyrics at the end of the prompt. Given the length of these prompts, it was advantageous to have all the relevant instructions at the beginning of the prompt in order to prevent/minimise any confusion for the model.
- We observed that two prompts from the same method can be different. While some prompts remain the same for both prompt contexts (e.g. Semi-Closed), some have had to be adapted in order for ChatGPT to generate the desired output. This is especially true for the Open prompts, whereby the Lyrics context comprised the additional instruction: ‘Your answer must be based on your understanding of the text’. We found out that, by including this sentence, we would minimise the risk of ChatGPT relying on knowledge it had acquired during training and force it to assign a topic label based on its own interpretation of the lyrics. Additionally, for the Knowledge (Open) prompt, we found that a single sentence sufficed for the model to yield the required output; by contrast, the Lyrics (Open) prompt comprised more details and some few-shot examples in order for the model to understand what was required of it.
- In the case of some prompts, there was a fine line between providing enough context while keeping instructions simple enough to avoid confusion. Below is an example of a ‘wrong’ prompt (Prompt 1) compared to the successful prompt (Prompt 2) that we used for Knowledge (Closed):

Prompt 1: **For each track and artist**, you must pick one label from this list **to describe the topic of the track**: Love, Friendship, Death, Depression, Sex, Heartache, Loneliness, Cheating, Religion and Drugs. You cannot pick any other label of your choice. For example: **All I Can Do** - 'Love'

Prompt 2: For each song, you must pick one label from this list: Love, Friendship, Death, Depression, Sex, Heartache, Loneliness, Cheating, Religion and Drugs. You cannot pick any other label of your choice. For example: 'Love'

As Prompt 1 was the first attempt of our experimentation process, we decided to include as much detail as possible by introducing the data ('for each track and artist'), defining the task ('to describe the topic of the track') and including the title of the first song from our dataset ('All I Can Do'). The desired output of the model would have provided the title of each song and its respective predicted topic label as illustrated by the one-shot example. The actual output was far from what was expected for two reasons: first, the song that was used in the prompt ('All I Can Do' - Dolly Parton) was skipped entirely by the model; second, the majority of the predicted labels were not from the pre-defined list of labels. Instead, labels such as 'Existentialism', 'Self-Reflection', 'Change' and 'Secrecy' were assigned.

As a result of this, we discovered that removing the 'surplus' instructions in the prompt (highlighted in bold) would result in our desired output. Although the reason for ChatGPT's erratic behaviour when presented with Prompt 1 is unclear, observations from this experiment gave us some meaningful information about the ideal level of context for the remaining prompts.

4.2.2 Final selection of prompts

Following our experimentation process, we established the following prompts for each method (Closed, Semi-Closed, Open) and context (Knowledge and Lyrics):

- **Knowledge (Closed):** [List of songs] For each song, you must pick one label from this list: Love, Friendship, Death, Depression, Sex, Heartache, Loneliness, Cheating, Religion and Drugs. You cannot pick any other label of your choice. For example: 'Love'
- **Knowledge (Semi-Closed):** [List of songs] For each song, you have two choices: The first choice is to choose a label from the following list : Love, Friendship, Death, Depression, Sex, Heartache, Loneliness, Cheating, Religion and Drugs - for example: 'Love'. If you don't agree with anything in the given list, pick any label of your choice - for example: 'Peace'
- **Knowledge (Open):** [List of songs] For each song, assign one topic label of your choice. For example: 'Love'
- **Lyrics (Closed):** [List of songs] Given the following song titles and their lyrics, extract only one topic from the following list to describe the song: Love, Friendship, Death, Depression, Sex, Heartache, Loneliness, Cheating, Religion and Drugs. Your answer must be in the following format: Song Title - Artist : Topic.
- **Lyrics (Semi-Closed):** [List of songs] For the following song title and lyrics, you have two choices: The first choice is to choose a label from the following list : Love, Friendship, Death, Depression, Sex, Heartache, Loneliness, Cheating, Religion and Drugs - for example: 'Love'. If you don't agree with anything in the given list, pick any label of your choice - for example: 'Peace'

- **Lyrics (Open):** [List of songs] Given the following song titles and lyrics, assign one topic label of your choice. For example: ‘Anger’, ‘Peace’ or ‘Reflection’. Your answer must be based on your understanding of the text.

4.3 Evaluation

4.3.1 Precision, Recall and F1

As a first step in our evaluation, we calculated the Precision, Recall and F1 score to observe the general performance of ChatGPT 3.5 and ChatGPT 4. Additionally, we carried out an evaluation for each class (i.e. the topic label). In our Related Work section, we observed how Papazoglou and Gaizauskas (2021) and Choi et al. (2014) make use of some of the same labels as in their dataset: namely ‘Heartache’, ‘Cheating’, and ‘Loneliness’. Running these evaluations would therefore enable us to gain some insights into how a transformer-based model such as ChatGPT performs compared to traditional models.

4.3.2 Calculating semantic similarity using cosine

As we would not be able to run standard evaluation metrics on the Semi-Closed and Open methods, we decided to use cosine similarity to quantify the semantic similarity between the Songfacts label and the predicted label generated by ChatGPT.

The formula and interpretation for Cosine similarity is as follows:

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

Figure 4.2: Cosine similarity formula

- **Vectors \mathbf{A} and \mathbf{B} :** The vectors \mathbf{A} and \mathbf{B} represent the two objects (in our case, the predicted and Songfacts labels) that we want to compare for similarity.
- **Dot product ($\mathbf{A} \cdot \mathbf{B}$):** The dot product of vectors \mathbf{A} and \mathbf{B} is a measure of how much they align with each other in the multi-dimensional space. It is the sum of the products of the corresponding elements in the two vectors.
- **Magnitudes of vectors ($\|\mathbf{A}\|$ and $\|\mathbf{B}\|$):** The magnitudes of vectors \mathbf{A} and \mathbf{B} represent their lengths in the multi-dimensional space (as established by Euclidean norms). These are calculated by taking the square root of the sum of the squares of all the elements in the vector.

To carry out these calculations, we employed Spacy (<https://spacy.io/>), an open-source library widely used in the NLP community. As it is designed for production usage, there are several advantages of using Spacy instead of other libraries: it can be used for an extensive range of tasks, is regularly evaluated and updated and comprises pre-trained word vectors. The latter point is especially relevant to us, as this will save time and resources instead of locating and loading our own embedding model.

Typically, Spacy’s models are available in a range of sizes: small (‘sm’), medium (‘md’) and large (‘lg’). Although it is possible to calculate similarity by using the ‘sm’ model, it only includes context-sensitive tensors, meaning that individual words/tokens will not have an assigned vector. In order to obtain the best results we therefore opted to use the ‘lg’ model, which contains 514,000 unique vectors.

4.3.3 Similarity thresholds

After obtaining a similarity score for each song, we grouped the songs under the following four similarity threshold categories:

- **Full match:** whereby the Songfacts label and predicted label are identical and obtain a similarity score of 1.
- **Almost match:** whereby the Songfacts label and predicted label obtain a similarity score of 0.81-1.
- **Substantial match:** whereby the Songfacts label and predicted label obtain a similarity score of 0.61-0.80.
- **Different interpretation:** whereby the Songfacts label and predicted label obtain a similarity score of 0.60 or under.

Assigning each song to one of these categories would facilitate our analysis and comparison of all three methods: Closed, Semi-Closed and Open.

Chapter 5

Results & Analysis

This chapter presents the results of our experiments: we begin with an overview of the main results, followed by a more in-depth analysis of pre-/post-2021 songs and an evaluation of the results per topic label.

5.1 Overview of main results

	Perfect match	Almost match	Substantial	Different interpretation
Knowledge (Closed)	47	0	3	50
Knowledge (Semi-Closed)	16	0	10	74
Knowledge (Open)	6	0	7	87
Lyrics (Closed)	54	0	6	40
Lyrics (Semi-Closed)	49	0	9	42
Lyrics (Open)	14	4	13	69

(a) Semantic similarity thresholds for ChatGPT3.5

	Perfect match	Almost match	Substantial	Different interpretation
Knowledge (Closed)	49	0	3	48
Knowledge (Semi-Closed)	47	0	3	50
Knowledge (Open)	6	0	9	85
Lyrics (Closed)	54	0	4	42
Lyrics (Semi-Closed)	45	0	5	50
Lyrics (Open)	5	2	7	86

(b) Semantic similarity thresholds for ChatGPT4

Table 5.1: Overview of main results for ChatGPT 3.5 and ChatGPT 4

This table presents the main results of the 6 methods for ChatGPT3.5 and ChatGPT4. To obtain these results, we calculated the cosine similarity between the predicted topic label and the gold standard topic label acquired by Songfacts for each song. Following this, we established four semantic similarity threshold groups: Perfect Match, Almost Match, Substantial Match and Different Interpretation. We then assigned each song to its corresponding semantic similarity threshold. (5.1).

Overall, the distribution of the songs seems to be relatively equal with an approximately 50/50 split between **Perfect Match** and **Different Interpretation**. The only exception for this is Knowledge (Semi-Closed) - ChatGPT3.5 and the Open methods. For example: for Knowledge (Open) - ChatGPT 4, the Perfect Match score is 6 and the Different Interpretation score is 85, which is a difference of 79 songs.

By looking at the **Perfect Match** scores, Lyrics (Closed) for both ChatGPT 3.5 and ChatGPT4 achieved the highest number of 54. For both versions, the lowest score was obtained by Knowledge (Open) with a score of 6. Out of the Perfect Match output for the Semi-Closed approach, all results were similar (49, 47 and 45) except for Knowledge (Semi-Closed) - ChatGPT 3.5, which achieved a much lower score of 16.

For both versions, Lyrics (Open) was the only time when a song was assigned an **Almost Match** label: 4 for ChatGPT3.5 and 2 for ChatGPT4. We will expand on these specific songs in section 5.1.3.

For both ChatGPT versions, there were at least 3 songs which were assigned the **Substantial Match** label. Given the limited scope of this project, we will not expand on songs which belong to the Substantial Match similarity threshold.

5.1.1 Overlapping songs (Closed Method)

In this table we present the number of overlapping songs with a Perfect Match score between combinations of prompt category and ChatGPT version. For example: for ChatGPT 3.5, there are 38 songs with a Perfect Match score which resulted from both the Knowledge and Lyrics prompt.

Prompt/ChatGPT version 1	Prompt/ChatGPT version 2	Overlap of Correctly Classified Songs
Knowledge 3.5	Lyrics 3.5	38
Knowledge 3.5	Knowledge 4	38
Knowledge 3.5	Lyrics 4	32
Lyrics 3.5	Knowledge 4	38
Lyrics 3.5	Lyrics 4	45
Knowledge 4	Lyrics 4	38
ALL		29

Table 5.2: Overlap of correctly classified songs for different prompt category/ChatGPT version combinations

This table indicates that a solid number of songs - 29 to 45 are correctly classified across different prompt category/ChatGPT version combinations. Most importantly, there are 29 songs with a Perfect Match score across the board. This implies that certain characteristics of a song could have impact on the results. We will elaborate on this in the following section.

5.1.2 Closed Method

In this section, we will be focusing on the results of Knowledge (Closed) and Lyrics (Closed) for both ChatGPT models.

As we can observe in 5.1, the Closed method has obtained the highest amount of occurrences of Perfect Match for both ChatGPT versions and prompt categories: ChatGPT 3.5 - Knowledge (Closed) with 47, ChatGPT 3.5 - Lyrics (Closed) with 54, ChatGPT 4 - Knowledge (Closed) with 49 and ChatGPT 4 - Lyrics (Closed) with 54.

The overlapping songs between prompt categories for ChatGPT3.5 and ChatGPT4 can be found in A.1 and A.2. Table A.1 contains the Perfect Match songs which overlap between Knowledge (Closed) and Lyrics (Closed) for ChatGPT3.5; Table A.2 contains the Perfect Match songs which overlap between Knowledge (Closed) and Lyrics (Closed) for ChatGPT4. Each table has 29 songs in bold - these are the songs which are featured in both tables.

Table A.5b contains the following statistics for these songs: the 5 most frequent words in the song, the total word count and the unique word count. Additionally, there are two more columns: ‘Topic word count (text)’, which indicates how many words of the 5 most

frequent words feature the topic label; and ‘Topic word count (title)’, which indicates how many words of the song title feature the topic label.

These statistics are important because we can gain insights into which characteristics of a certain song could lead to it being assigned a Perfect Match score. For example, if a song in the Friendship category has frequent instances of the word ‘friends’, then both models would have assigned it the label ‘Friendship’, and could therefore be categorised as Perfect Match. This has proved to be the case for the ‘Best Friends’ by The Weeknd, ‘Jesus Lord’ by Kanye West, ‘Between the Cheats’ by Amy Winehouse and ‘Never Felt So Alone’ by Billie Eilish. In A.5b we can see that all of these 4 songs have their respective topic word (Friendship, Religion, Cheating and Loneliness) featured in both the 5 most frequent words and in the title. From these observations, we could make a preliminary conclusion that the overlap between lyrics and the title of the song could increase the likelihood of accurate classification.

5.1.3 Semi-Closed Method

In this section, we will be focusing on the results of Knowledge (Semi-Closed) and Lyrics (Semi-Closed) for both ChatGPT versions.

For the Semi-Closed method, the model is given a choice between choosing labels from a set list or free choice to assign its own label. As we can see in 5.1, almost all of the Semi-Closed methods yielded between 45 and 49 (out of 100) Perfect Match labels. The only exception was GPT3.5 - Knowledge (Semi-Closed), which resulted in 16 instances of Perfect Match, compared to 47 from GPT4 - Knowledge (Semi-Closed). A potential reason for this could be that ChatGPT4 has been trained on a far greater amount of internet data than ChatGPT3.5; therefore, it is possible that ChatGPT4 would have been exposed to a higher volume of information regarding the topic of the songs in question, which in turn would have enabled it to generate more accurate predictions.

The overlapping songs between prompt categories for ChatGPT3.5 and ChatGPT4 can be found in A.3 and A.4. Table A.3 contains the Perfect Match songs which overlap between Knowledge (Semi-Closed) and Lyrics (Semi-Closed) for ChatGPT3.5; Table A.4 contains the Perfect Match songs which overlap between Knowledge (Semi-Closed) and Lyrics (Semi-Closed) for ChatGPT4. Each table has 29 songs highlighted in bold: these are the songs which are featured in both tables and have therefore achieved a Perfect Match in both prompt categories and ChatGPT versions. Examples of these songs are: ‘Love Song’ by Lana del Rey, ‘Crazy Little Thing Called Love’ by Queen and ‘All I Can Do’ by Dolly Parton. This is potentially because these songs date from pre-2021 and are associated to popular artists that have been active for a considerable amount of years before 2021. In regards to the Knowledge prompt category, it is likely that both ChatGPT versions would have already been familiar enough with the songs to make an accurate prediction.

The only overlapping song which was post-2021 was ‘Cuff It’ by Beyonce. In the case of the Knowledge prompt category, although the models had not been exposed to the lyrics, the title contained enough information for the models to assign the correct topic label. As a result, we can make an initial claim that ChatGPT does indeed rely on the title of the song to inform its predictions. In the case of the Lyrics prompt, the message conveyed in the song lyrics was strong enough for the models to make an accurate prediction. These are similar observations which were also found in 5.1.2. Following these observations, we could also make a preliminary claim that the date in which a song is released is not as influential as the content and the title of the song itself. We will be further expanding on this latter point in Section 5.2.

5.1.4 Open Method

In this section, we will focus on the results of Knowledge (Open) and Lyrics (Open) for both models. As we observe in 5.1, the majority of the results are classified as Different Interpretation labels. Because ChatGPT is given freedom to choose its own topic labels, the resulting labels are very diverse compared to those from the Closed and Semi-Clothed methods.

By looking at 5.1, we can see that only time a song was assigned to the Almost Match category was for the Lyrics (Open) method: 4 instances for GPT3.5 and 2 instances for GPT4. We identified the songs as the following:

Track	Artist	Pre-2021	Songfacts label	Predicted label
Back to Black	Amy Winehouse	Yes	Heartache	Heartbreak
Don't You	Taylor Swift	No	Heartache	Heartbreak
Let Somebody Go	Coldplay	No	Heartache	Heartbreak
Hits Different	Taylor Swift	No	Heartache	Heartbreak

Table 5.3: Almost Match songs for Lyrics (Open) - GPT3.5

??

Track	Artist	Pre-2021	Songfacts label	Predicted label
Back to Black	Amy Winehouse	Yes	Heartache	Heartbreak
Hits Different	Taylor Swift	No	Heartache	Heartbreak

Table 5.4: Almost Match songs for Lyrics (Open) - GPT4

A very interesting parallel can be drawn from these songs: they all belong to the Songfacts topic label 'Heartbreak', and they have all been labelled by the models as 'Heartache'. For this reason, all of these instances were assigned a similarity score of 0.85, and therefore grouped in the Almost Match category. For the Closed and Semi-Open methods, the model is presented with the term 'Heartache' in the prompt. This is not the case for the Open method, as the models are given free choice to assign their own topic label. It is therefore interesting to see that, when the term 'Heartache' is not included in the prompt, the models opt for the label 'Heartbreak' instead. In cases like this, some form of post-categorisation by a human evaluator would probably be required.

For both ChatGPT versions, there are 6 songs for Knowledge (Open) which have been assigned to the Perfect Match category. These songs can be found in the following tables:

Track	Artist	Pre-2021	Songfacts label	Predicted label
Love Song	Lana del Rey	Yes	Love	Love
Bobby Jean	Bruce Springsteen	Yes	Friendship	Friendship
Grigio Girls	Lady Gaga	Yes	Friendship	Friendship
Best Friends	The Weeknd	Yes	Friendship	Friendship
Bad	Frank Ocean	Yes	Drugs	Drugs
Eleanor Rigby	Beatles	Yes	Loneliness	Loneliness

Table 5.5: Perfect Match songs for Knowledge (Open) - GPT3.5

The only overlapping song from these tables is 'Love Song' by Lana del Rey. As we already discovered in 5.1.2 and 5.1.3, if the title of the song is descriptive and reflects the message conveyed in the song lyrics, there is a higher chance of the model assigning the correct topic label. This is also the case for songs which are pre-2021, as the model is likely

Track	Artist	Pre-2021	Songfacts label	Predicted label
Love Song	Lana del Rey	Yes	Love	Love
Maureen	Sade	Yes	Friendship	Friendship
Lithium	Nirvana	Yes	Depression	Depression
Muddy Feet	Miley Cyrus	No	Cheating	Cheating
Praise God	Kanye West	No	Religion	Religion
Never Felt so Alone	Billie Eilish	No	Loneliness	Loneliness

Table 5.6: Perfect Match songs for Knowledge (Open) - GPT4

to have already been exposed to data associated to the song (or even the song itself) during pre-training. This case can be observed in 5.5, which includes only songs from pre-2021. On the other hand, 5.6 features 3 songs which are post-2021: ‘Muddy Feet’ by Miley Cyrus, ‘Praise God’ by Kanye West and ‘Never Felt so Alone’ by Billie Eilish. Out of these songs, we can argue that the titles ‘Praise God’ and ‘Never Felt So Alone’ are relatively self-explanatory and contain enough information for the models to assign the ‘correct’ topic label, regardless of the date of their release. Conversely, the title of ‘Muddy Feet’ is not as easily interpretable, yet the model was still able to generate an accurate prediction. One possible reason for this is that Miley Cyrus might have already released a number of songs which talk about infidelity or broken relationships before 2021. Therefore, the model’s prediction could be a result of it making an informed guess based on the artist’s past discography, of which it has pre-existing knowledge.

5.2 Results of Pre- and Post-2021 songs

Prompt Category	ChatGPT3.5		ChatGPT4	
	Pre-2021	Post-2021	Pre-2021	Post-2021
Knowledge (Closed)	29(*)	21	29	23
Lyrics (Closed)	29	31	28	28

(a) Distribution of pre- and post-2021 songs - ChatGPT3.5 and ChatGPT4 (Closed). The results are an aggregation of songs which are assigned Perfect Match, Almost Match or Substantial Match.

Prompt Category	ChatGPT3.5		ChatGPT4	
	Pre-2021	Post-2021	Pre-2021	Post-2021
Knowledge (S-C)	14	12	26	24
Lyrics (S-C)	28	30	26	26

(b) Distribution of pre- and post-2021 songs - ChatGPT3.5 and ChatGPT4 (Semi-Closed). The results are an aggregation of songs which are assigned Perfect Match, Almost Match or Substantial Match.

Table 5.7: Distribution of pre- and post-2021 songs - ChatGPT3.5 and ChatGPT4

Table 5.7 contains the aggregated number of songs which are Perfect Match, Almost Match or Substantial Match which are either pre-2021 or post-2021. For example, 29(*) in 5.7 shows that ChatGPT3.5 assigned the label Perfect Match, Almost Match or Substantial Match to 29 songs which are pre-2021.

This table is divided in two sub-tables: (a) represents the results for the Closed method and (b) represents the results for the Semi-Closed method. Because the Open method has obtained a much lower number of Perfect Match, Almost Match or Substantial Match scores (see 5.1), we did not include the results of the Open method in this section.

It is important to assess the differentiation between pre-2021 and post-2021 songs because we wanted to observe if the performance of either ChatGPT versions is affected by the date of the song. A first preliminary hypothesis was that, in regards to the Knowledge prompt category, post-2021 songs would obtain a lower amount of Perfect Match/ Almost Match/ Substantial Match due to the fact that the songs would not have been present in ChatGPT's training data. However, we observed a relatively equal balance between post-2021 and pre-2021 songs for both ChatGPT versions and across prompt categories and methods. As a result of this, there is no tangible difference in performance between pre-2021 and post-2021 songs.

A second preliminary hypothesis was that post-2021 songs would be better classified when using the Lyrics prompt over the Knowledge prompt. This was proven right in a number of cases: ChatGPT3.5 - Closed (21 vs 31); ChatGPT4 - Closed (23 vs 28), ChatGPT3.5 - Semi-Closed (12 vs 30) and ChatGPT4 - Semi-Closed (24 vs 26).

In this section, we explained how post-2021 songs achieved comparable results to pre-2021 songs, regardless of whether or not ChatGPT has been exposed to certain songs during pre-training. As a result, we can make a preliminary conclusion that the year in which a song was released may not necessarily be an influencing factor in ChatGPT's decision making. In regards to the Knowledge prompt, it may rely on the content and the title of the song itself to make its predictions, as already observed in sections 5.1.2 and 5.1.3. When comparing the Knowledge prompt to the Lyrics prompt, the Lyrics prompt performs better overall compared to the Knowledge prompt, as also observed in Table 5.1.

5.3 Results per topic label

In this section, we focus on the topic labels themselves to assess whether some topics are easier to identify and classify compared to others. In order to generate accurate observations, we only take into consideration the results from the Closed method. We will first look into the Precision, Recall and F1 score of each class across both ChatGPT versions and prompt categories; following this, we will present the results as confusion matrices in order to observe which classes get easily confused and why.

5.3.1 Results per class

Class	V(P)	Precision	Recall	F1 Score	V(P)	Precision	Recall	F1 Score
Cheating	3.5 (K)	0.8	0.4	0.53	4 (K)	1.0	0.4	0.57
Death	3.5 (K)	0.67	0.6	0.63	4 (K)	0.88	0.7	0.78
Depression	3.5 (K)	0.75	0.3	0.43	4 (K)	0.38	0.5	0.43
Drugs	3.5 (K)	0.44	0.4	0.42	4 (K)	1.0	0.3	0.46
Friendship	3.5 (K)	0.45	0.5	0.48	4 (K)	0.62	0.5	0.56
Heartache	3.5 (K)	0.19	0.3	0.23	4 (K)	0.17	0.2	0.18
Loneliness	3.5 (K)	0.5	0.7	0.58	4 (K)	0.88	0.7	0.78
Love	3.5 (K)	0.29	0.5	0.37	4 (K)	0.2	0.6	0.3
Religion	3.5 (K)	0.75	0.6	0.67	4 (K)	0.64	0.7	0.67
Sex	3.5 (K)	0.57	0.4	0.47	4 (K)	1.0	0.3	0.46
Cheating	3.5 (L)	0.86	0.6	0.71	4 (L)	1.0	0.6	0.75
Death	3.5 (L)	1.0	0.6	0.75	4 (L)	1.0	0.5	0.67
Depression	3.5 (L)	0.38	0.3	0.33	4 (L)	0.38	0.5	0.43
Drugs	3.5 (L)	0.57	0.4	0.47	4 (L)	0.75	0.3	0.43
Friendship	3.5 (L)	0.57	0.4	0.47	4 (L)	0.83	0.5	0.62
Heartache	3.5 (L)	0.24	0.5	0.32	4 (L)	0.25	0.4	0.31
Loneliness	3.5 (L)	0.56	0.5	0.53	4 (L)	0.38	0.5	0.43
Love	3.5 (L)	0.5	0.9	0.64	4 (L)	0.43	0.9	0.58
Religion	3.5 (L)	0.67	0.6	0.63	4 (L)	0.8	0.8	0.8
Sex	3.5 (L)	0.75	0.6	0.67	4 (L)	0.67	0.4	0.5

Table 5.8: Results per class for ChatGPT3.5 and 4

This table presents the Precision, Recall and F1 scores for each class and ChatGPT version/prompt category combination. The latter is represented in the table as V(P), which are the initials of ‘Version (Prompt)’.

We will now list the two classes which obtained the highest and lowest F1 scores for each ChatGPT version and prompt category.

For ChatGPT3.5 (Knowledge): **Death** (0.63) and **Religion** (0.67) were the two highest-scoring classes. The two lowest-scoring classes were **Heartache** (0.23) and **Love** (0.37). For ChatGPT3.5 (Lyrics): **Death** (0.75) and **Cheating** (0.71) were the two highest-scoring classes. The two lowest-scoring classes were **Heartache** (0.32) and **Depression** (0.33).

For ChatGPT4 (Knowledge), the two highest-scoring classes were **Death** (0.78) and **Loneliness** (0.78). The two lowest-scoring classes were **Love** (0.30) and **Heartache** (0.18). For ChatGPT4 (Lyrics), the two highest-scoring classes were **Religion** (0.80) and **Cheating** (0.75). The two lowest-scoring classes were **Heartache** (0.31) and **Depression** (0.43) and **Loneliness** (0.43).

To summarise, **Heartache** was the lowest-scoring class for all four ChatGPT versions and prompt category. **Death** was the class which was the most frequent in the two highest-performing classes: (3 out of 4 times). For ChatGPT3.5 Knowledge and ChatGPT4 Knowledge, **Heartache** and **Religion** were the two lowest-scoring classes.

5.3.2 Confusion Matrices for the Closed Method

In this section, we present our results in four confusion matrices, one for each ChatGPT version (3.5 and 4) and prompt category (Knowledge and Lyrics). Following this, we highlight the instances in which a particular topic label is frequently misclassified and identify the cause by carrying out an analysis.

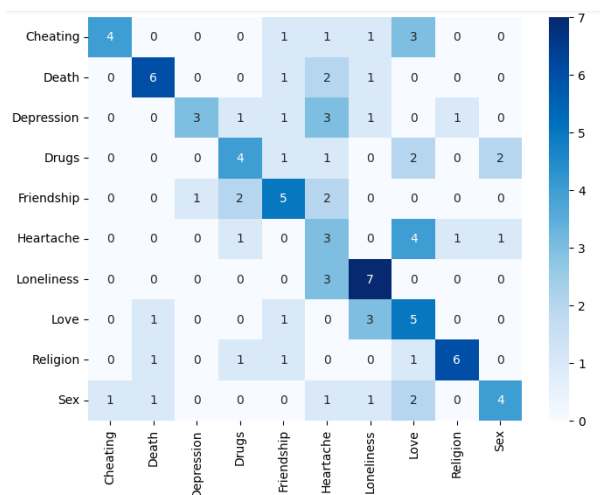


Figure 5.1: GPT3.5 Confusion Matrix: Knowledge

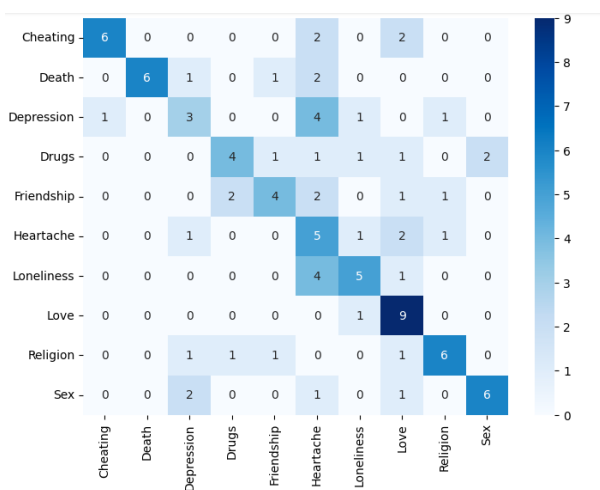


Figure 5.2: GPT3.5 Confusion Matrix: Lyrics

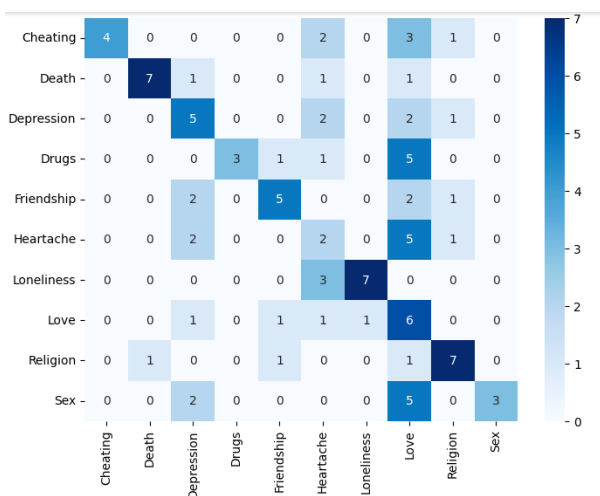


Figure 5.3: GPT4 Confusion Matrix: Knowledge

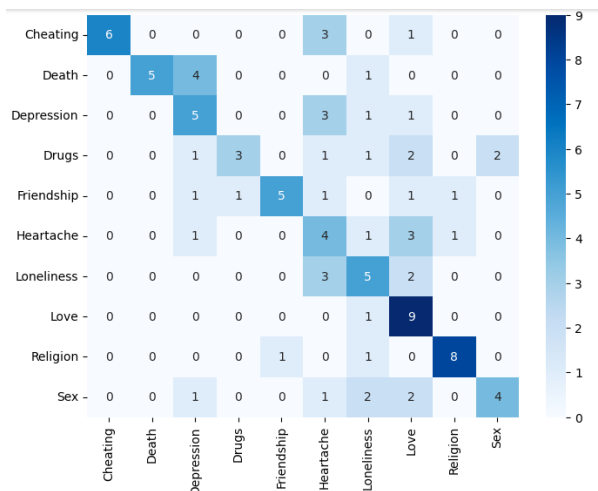


Figure 5.4: GPT4 Confusion Matrix: Lyrics

Overall, **Love** is the topic category that is the most confused with other categories. ChatGPT4 (Knowledge) appears to have the highest instances of confusion for Love: for example, with Drugs (5), Loneliness (5) and Sex (5).

Heartache is also often misclassified across all ChatGPT versions and prompt categories. The classes with which it is most confused are Loneliness, Cheating and Depression.

5.3.3 Analysis of the Confusion Matrices for the Closed Method

In this section we will explore the results generated in Table 5.8 and Figures 5.1 - 5.4.

Perhaps the most interesting insight that was uncovered from Table 5.8 was that **Heartache** was the lowest-scoring class for all four ChatGPT versions and prompt categories. In our Related Work section, we uncovered how Papazoglou and Gaizauskas (2021) and Choi et al. (2014) also found out that Heartache - along with Cheating and Loneliness - was the most frequently confused class in their experiments. We will now present our main findings of the confusion matrices in more detail.

5.3.4 ChatGPT3.5 - Knowledge

In this confusion matrix, **Heartache** was labelled as **Love** 4 times. These mislabelled songs are:

Track	Act	Pre-2021	Key lyrics
A Different Corner	George Michael	Yes	I'd say love was a magical flame / I'd say love would keep us from pain
I Can't Quit You Baby	Led Zeppelin	Yes	Said you know I love you baby / My love for you I could never hide
All of Me	Frank Sinatra	Yes	You took my kisses and you took my love / Am I to be just the remnant of a one-sided love affair
Don't You	Taylor Swift	No	You don't know how much I feel I love you still

Table 5.9: Heartache songs mislabelled as 'Love' - ChatGPT3.5 (Knowledge)

From this selection, we notice two main insights: 3 out of 4 songs are pre-2021, and the lyrics of all the songs mention the term 'love' at least once. In regards to the post-2021 songs, it is possible that ChatGPT will already have been exposed to the lyrics during pre-training;

therefore, we could argue that ChatGPT assigned the label ‘Love’ based on the frequency of the term ‘love’ in these songs. Although ChatGPT is able to process and understand the entire context of a given input/text, we could argue that the limitations of the Closed method might have pushed ChatGPT into not looking further than the frequency count of a given word.

The only post-2021 song was ‘Don’t You’ by Taylor Swift. Although ChatGPT won’t have been exposed to its lyrics during training, it is possible that it has based its prediction on previous knowledge of the artist given that Taylor Swift is known for writing about relationships (Knot).

5.3.5 Chat GPT3.5 - Lyrics

In this confusion matrix, **Depression** is labelled as **Heartache** 4 times. These mislabelled songs are:

Track	Act	Pre-2021	Key lyrics
Cleaning my Gun	Chris Cornell	Yes	Lovers’ game / Somehow you decided you would find another flame
Borderline	Ed Sheeran	No	N/A
Cry Your Heart Out	Adele	No	Your love is useless without it / Cry your heart out
Forever Winter	Taylor Swift	No	I’d say I love you even at your darkest / And please don’t go /... /I’d fall to pieces on the floor

Table 5.10: Depression songs labelled as ‘Heartache’ - ChatGPT3.5 (Lyrics)

From these songs, we could not find lyric excerpts in ‘Borderline’ by Ed Sheeran which could indicate why it was classified as ‘Heartache’ instead of ‘Depression’. Conversely, the lyrics of the remaining 3 songs may allude to feelings of heartache as a result of a failed relationship: perhaps it is the presence of the term ‘love’ along with the overall negative sentiment of the song which might have led to the prediction of ‘Heartache’. This is particularly the case for ‘Cleaning my Gun’ by Chris Cornell, which is the only pre-2021 song.

5.3.6 ChatGPT4 - Knowledge

A large amount of instances are misclassified as Love: for example Heartache (5), Sex (5) and Drugs (5). While the former two labels are understandable given their semantic proximity to the term ‘Love’, it is interesting that songs which are originally labelled as Drugs are mislabelled as Love. Because of this, we are presenting the 5 mislabelled song together with an excerpt of their description from Songfacts. On the Songfacts website, each song has its own page containing information about its release date as well as interviews and press releases which talk about the conception or the message behind the song. For this section, we examined the Songfacts information page for each song and picked a relevant excerpt which could provide a reason for the misclassification.

Track	Act	Pre-2021	Songfacts quotes
A Baltimore Love Thing	50 Cent	Yes	'This song also portrays a love relationship with a girl similar to a relationship that someone would have with heroin' (Son (a))
Beauty and the Beast	David Bowie	Yes	'Ode to Bowie's love/hate relationship with cocaine' (Son (b))
Brown Sugar	D' Angelo	Yes	'Not an ode to a dark-skinned woman, but it's a love song to marijuana' (Son (c))
Tippa My Tongue	Red Hot Chili Peppers	No	'Tippy love song with ambiguous lyrics and drug references' (Son (d))
Bad Habits	Ed Sheeran	No	'People see me as the acoustic singer-songwriter who does ballads..' (Son (e))

Table 5.11: Drugs songs which were mislabelled as 'Love' - ChatGPT4 (Knowledge)

The first 4 Songfacts quotes explain how the lyrics either use drug references as a metaphor for relationships or chronicle the artists' own 'love' affair with drugs. The intertwining of these two concepts - love and drugs - are therefore likely the cause for this misclassification.

In regards to 'Bad Habits' by Ed Sheeran, Songfacts' description states how Sheeran's desire to write different material from what is normally expected of him: 'People see me as the acoustic singer-songwriter who does ballads and there was just a lot of that ... So I wanted to go in the studio and make something that was totally different.' Because 'Bad Habits' was released post-2021, it is possible that ChatGPT4 would have based its prediction on pre-existing knowledge of the artists' discography.

5.3.7 ChatGPT4 - Lyrics

In the confusion matrix, **Loneliness**, **Depression** and **Cheating** are all classified as **Heartache** 3 times.

Track	Act	Pre-2021	Key lyrics
Biggest Mistake	Rolling Stones	Yes	'But if love comes again, I'll be really surprised ... Cause I think I've just made the biggest mistake of my life
Heartbreak Hotel	Elvis Presley	Yes	I'm so lonely I'll be so lonely, I could die ... You still can find some room For broken hearted lovers
Old Memories	Alicia Keys	No	Old love songs They don't ever end Just when you think that you moved on They remind you, you ain't over it

Table 5.12: Loneliness songs which were mislabelled as 'Heartache' - ChatGPT4 (Lyrics)

Track	Act	Pre-2021	Key lyrics
Cleaning My Gun	Chris Cornell	Yes	Mama always told me love would save me from myself Daddy always said that love would take me straight to hell
Boats	Ed Sheeran	No	The more that I love the less that I feel The times that I jumped never were real They say that all scars will heal but I know Maybe I won't
Cry Your Heart Out	Adele	No	Your love is useless without it Cry your heart out

Table 5.13: Depression songs which were mislabelled as ‘Heartache’ - ChatGPT4 (Lyrics)

Track	Act	Pre-2021	Key lyrics
Cold Shoulder	Adele	Yes	Do tell me why you waste our time When your heart ain't in it, and you're not satisfied
Muddy Feet	Miley Cyrus	No	Get the f*ck out of my life with that shit You smell like perfume that I didn't purchase Now I know why you've been closing the curtains, ah
One Right Now	Post Malone	No	You say you love me, but I don't care That I broke my hand on the same wall That you told me that he f*cked you on

Table 5.14: Cheating songs which were mislabelled as ‘Heartache’ - ChatGPT4 (Lyrics)

By looking at the lyrics of all of these instances, we can detect references to events or feelings which allude to heartache. Furthermore, we can argue that the three labels Loneliness, Depression and Cheating carry a similar degree of negative sentiment to Heartache. These are the potential factors for the mislabelling of these instances, which are also in line with the findings of Papazoglou and Gaizauskas (2021) and Choi et al. (2014) whereby their models confuse labels which share the same sentiment (2).

In this section, we focused on the performance of the topic labels: Cheating, Death, Depression, Drugs, Friendship, Heartache, Loneliness, Love, Religion and Sex. The aim of this was to identify which topic labels were the most frequently confused by ChatGPT, and why.

First, we calculated the Precision, Recall and F1 score of each class. From this, we discovered that Heartache was the lowest-scoring class for all four ChatGPT versions and prompt categories. Death was the class which was the most frequent in the two highest-performing classes: (3 out of 4 times).

To determine which topic labels were most confused with other labels, we presented the 4 confusion matrices for each ChatGPT version and prompt category. Following this, we carried out an analysis of the most frequent instances of misclassification. For **ChatGPT3.5 - Knowledge**, Heartache was labelled as Love 4 times; for **ChatGPT3.5 - Lyrics**, Depression was labelled as Heartache 4 times; for **ChatGPT4 - Knowledge**, Drugs was labelled as Love 5 times; for **ChatGPT4 - Lyrics** Loneliness, Depression and Cheating were all classified as Heartache 3 times.

There are principally two reasons for these misclassifications: one is due to the presence of a certain word in the lyrics which could lead to confusion for the model; for example, for ChatGPT3.5 - Knowledge, the presence of the word ‘love’ in some songs has led to them being assigned the Love label instead of Heartache. The second reason is that some topic labels are confused with others because they have the same sentiment. For example, for ChatGPT4 - Lyrics, the meaning of the labels Loneliness, Depressing and Cheating share the same negative sentiment as Heartache, which could lead the model to misclassify these

songs.

Chapter 6

Conclusion & Discussion

In this section we present our discussion, conclusion and recommendations for future research.

6.1 Discussion

We will begin the discussion by answering our research questions:

- **RQ1. Can ChatGPT be used for topic extraction on songs, based on a set of predefined topics vs. a free choice of topics?**

To summarise, ChatGPT is able to perform topic extraction on lyrics to a certain extent. To achieve more accurate results, we discovered that it is best to use the combination of the Closed method and Lyrics prompt: in other words, forcing ChatGPT to make predictions by choosing from a set list of topic labels and including the lyrics of each song in the prompt. This conclusion is due to the fact that both ChatGPT versions (3.5 and 4) obtained the highest instances of Perfect Match scores (54 each) by using this combination.

Additionally, we gained insights into what characteristics of a song could help ChatGPT generate more accurate predictions. In 5.1.2 and 5.1.3, we observed some instances in which the topic label of a song was featured in both its title and in its 5 most frequently occurring words; therefore, it would possibly be more likely to achieve a Perfect Match score. However, table A.5b shows that around a third of songs (10 out of 29) which are awarded a Perfect Match score don't have their topic label featured in their title or their 5 most frequently occurring words: for example, 'Meaning of Life' by Kelly Clarkson and 'Candy Shop' by 50 Cent.

We also uncovered that ChatGPT encountered the same limitations as traditional models by confusing topic labels that are associated with negative sentiment: Loneliness, Heartache and Cheating (5.3.3). Therefore, we could argue that ChatGPT may have some issues capturing certain nuances in song lyrics which could help it generate more accurate predictions. This could potentially also be due to the characteristics of lyrical songwriting, where the meaning and emotion could be more ambiguous compared to the type of data that ChatGPT has been trained on (e.g. factual content such as newspaper articles). This also opens discussion into the extent to which ChatGPT's performance is influenced by the topic labels themselves. In other words: our selection process for the 10 gold standard topic labels was focused on whether or not there were enough songs from pre-2021 and post-2021, without taking the sentiment or semantic properties of the label into consideration. If we had curated a different selection of topic labels as gold standard, there is a chance that the results regarding the best-performing ChatGPT version, method or prompt category would also be different.

In addition to this, we discovered some instances in which misclassification would occur as a result of a song featuring more than one topic. Therefore, even though the predicted label was different from the gold label obtained by Songfacts, it did not necessarily mean that it was incorrect (5.3.7). This opens up discussion into how or why ChatGPT chooses a certain topic label over another (even if they are both correct) and finding ways to improve and develop our dataset in order to further explore such questions.

By using the Semi-Closed and Open methods, we gained insights into how ChatGPT behaves when it is not restricted to choose from a set list of topic labels. As a whole, the results from both methods were not as accurate or consistent as the Closed method (5.1. The Open method only produced between 6 and 14 Perfect Match scores; conversely, the Semi-Closed method did produce comparable results to the Closed method by obtaining between 45 and 49 Perfect Match scores. However, it also scored only 16 Perfect Match scores for Knowledge (Semi-Closed), which might point to some inconsistency and lack of reliability in this method.

Despite this, by not limiting ChatGPT to follow a set number of rules, we were able to uncover insights into its behaviour and approach to making predictions. Although the results of the Semi-Closed and Open methods aren't as accurate as the Closed method, there are still interesting observations to be made about ChatGPT's behaviour. For example, the Open method revealed that when ChatGPT is presented with a post-2021 song without being given the lyrics in the prompt, it tends to make its prediction based on the title of the song (5.3.7). Therefore, if a song title is descriptive enough and is also consistent with the song lyrics, ChatGPT is more likely to assign a topic label that is identical or semantically similar to the Songfacts label. Additionally, we observed an instance in which both versions of ChatGPT assigned the topic label 'Heartbreak' to songs with the Songfacts label 'Heartache' (5.3 and 5.4). Though these instances were not classified as a Perfect Match, it can be argued that they have the exact same meaning as the Songfacts label. This encourages further questions into the extent to which the semantic distance between two words should determine whether a prediction is correct or not. The above point provides an opportunity for us to begin confronting limitations in our research. Due to the subjectivity surrounding natural language and the interpretation of song lyrics, our research could have benefited from involving human evaluation to validate or dispel our quantitative results. This is particularly the case for the Semi-Closed and Open methods, which are expected to predict (mostly) songs which are not from the set list of topics. In other words, just because a predicted label has been categorised as Different Interpretation (and is therefore semantically different) does not necessarily mean that it is not correct or relevant to the topic of the song in question.

The employment of the Semi-Closed and Open methods have given interesting insights into the behaviour of ChatGPT and opened up valuable discussion points. However, given the variable results, it is not clear how and if these methods would be immediately useful to XITE's Music team. In other words, a certain level of post-categorisation and human evaluation would be needed to optimise these methods. The same can also be said of the Closed method, despite the fact that it has achieved the highest amount of Perfect Match scores across these three methods. As the highest score achieved was of 'only' 54 Perfect Match scores, further research is needed to identify flaws and circumstances which impede the performance of ChatGPT. Nevertheless, ChatGPT's accessibility and breadth of knowledge can be advantageous to XITE's Music team for extracting information of a song based on its lyrics, to a certain extent.

- **RQ 1a) To what extent does the release date (pre- or post-2021) of the song**

affect ChatGPT ’s ability to do topic extraction on songs?

In regards to the release date of the song, we found that the results from the Closed method were not necessarily impacted by whether or not a song was released before or after 2021. This went against expectations that the majority of correct predictions would come from pre-2021 songs, given that ChatGPT has been trained on data which dates up to 2021 (5.2). In fact, using the Lyrics prompt caused post-2021 songs to perform as well or better than pre-2021 songs. This might be due to the fact that ChatGPT has not been exposed to these songs during pre-training; as a result, it did not have access to any information relating to the song other than the song lyrics themselves.

- **RQ 1b) How does including the lyrics in the prompt affect ChatGPT ’s ability to do topic extraction on songs?**

We observed that the inclusion of the lyrics in the prompt produced more accurate results than relying on ChatGPT’s pre-existing knowledge of a given song or artist (5.1). Because there are no available details about ChatGPT’s training data and process, we do not know the extent to which ChatGPT is familiar with a certain song. Including the lyrics in the prompt has shown to be especially effective for post-2021 songs, as ChatGPT has not been exposed to these songs during pre-training. Therefore, forcing ChatGPT to base its interpretation on a given set of lyrics minimises the risk of it making assumptions and, in turn, incorrect predictions.

- **RQ 1c) To what extent does the version of ChatGPT (3.5 or 4) affect its ability to perform prompt-guided topic extraction?**

As stated in our answer to RQ1, both ChatGPT versions (3.5 and 4) obtained the highest instances of Perfect Match scores (54 each) by using the combination of the Closed method and Lyrics prompt. In the case of this experiment, the actual difference between these two versions of ChatGPT is very minimal. However, using ChatGPT3.5 for prompt-guided topic extraction could more beneficial for the following reasons:

- It achieves comparable results to ChatGPT4, which is more advanced and has been reported to perform better than ChatGPT3.5 over other NLP tasks.
- It is free to use and does not have the same limited cap-per-usage that ChatGPT4 has.

Therefore, for the specific purpose of prompt-guided topic extraction, ChatGPT3.5 would be equally effective and more accessible than ChatGPT4.

6.2 Conclusion

This research focused on evaluating ChatGPT’s ability to perform prompt-guided topic extraction on song lyrics. Our aim is to provide recommendations to XITE’s Music team for extracting information of a song based on its lyrics. The solution must fulfil the following criteria: it must be scalable, robust and easy-to-use for the Music team. With this in mind and considering the recent developments of state-of-the-art Natural Language Processing (NLP) systems, we decided to use ChatGPT to carry out the task.

The motivation for this was twofold:

1. To provide XITE’s Music team with recommendations on how to extract information from song lyrics in an (semi-)automatic, scalable way.

2. To assess how the currently-available versions of ChatGPT (3.5 and 4) were able to perform topic extraction on song lyrics given different factors: the method (Closed, Semi-Closed, Open), the prompt category (Lyrics and Knowledge) and the year of release of each song (Pre-2021 and Post-2021).

The reason that we decided to use two prompt categories (Knowledge and Lyrics) and year thresholds (pre-2021 and post-2021) is because ChatGPT has been trained on large amounts of textual data obtained from the internet (e.g. Wikipedia articles, books and websites) up to September 2021. Because XITE’s database is regularly updated with new entries (i.e. songs which are released after 2021), we want to test ChatGPT’s ability to process lyrics of songs which it won’t have been exposed to during pre-training. Additionally, we wanted to assess whether including lyrics in the prompt would improve the performance of ChatGPT.

Our approach consisted of two steps: first, the selection and creation of our own labelled dataset by using the Songfacts database; second, the implementation of the experiments in addition to the comparison of results between ChatGPT3.5 and ChatGPT4.

The value of this research stems from the fact that - to the best of our knowledge - no previous scientific study has been carried out on ChatGPT’s ability to process and extract information from song lyrics. Furthermore, our findings could potentially optimise the workflow and metadata-labelling process of XITE’s Music team. Our findings led us to conclude that using ChatGPT3.5 and the Lyrics prompt is the best option to perform prompt-guided topic extraction. On a broader level, incorporating the lyrics in the prompt is more likely to produce accurate results (5.7). Furthermore, we observed that the year of release of the song does not necessarily affect ChatGPT’s performance and ability to make accurate predictions (5.2).

These findings also opened up discussion points where we challenged the reliability of certain methods (e.g. Semi-Closed) and the extent to which semantic and sentimental characteristics of the topic labels influence ChatGPT’s performance, rather than the song lyrics themselves. Additionally, we touched on the extent to which ChatGPT and the methods can be useful to XITE’s Music team.

6.3 Future work

Following our discussion and conclusion, we are now able to make recommendations for further research in the future.

In terms of our dataset, while only having 100 songs was conducive to carrying out a more in-depth evaluation and interpretation, it is not necessarily representative of the actual size of XITE’s database. Therefore, the results obtained in this research might not reflect those that would be obtained from processing lyrics for hundreds of thousands of songs. Additionally, XITE’s database also features songs in different languages, while we limited our data collection to songs that were originally written in English. While our dataset is already representative of different information categories (e.g. popularity, pre/post-2021 and lyrics), we recommend a larger scale study in the future by capturing more topics, song types and features.

For the purpose of this project, we limited our dataset to only one topic label per song, despite the fact that some songs were assigned multiple topic labels by Songfacts. In the future, we would include all topic labels of the song. In our work we found that certain labels such as Love, can be often mislabeled due to being broad. Also, although some of ChatGPT’s suggestions were semantically distant to the given label lists, qualitatively they were found to carry similar sentiment to the original label. Therefore, we suggest in the future the inclusion of a sentiment tag for each song to determine whether it carries positive, neutral or negative sentiment. The combination of a sentiment and topic label could help

add nuances to topics which are broad, such as ‘Love’. Applying these differentiations could help tie songs that are semantically different but similar in sentiment: for example, we could associate songs which are ‘Love (Negative)’ to other songs from ‘Heartache (Negative)’ or ‘Cheating (Negative)’. As a result of the inclusion of multiple topic labels and sentiment labels, XITE’s Music team would be presented with a diverse selection of songs which could all be used in the same themed playlist/channel.

In particular regard to the Semi-Closed and Open method, we would apply additional steps to assess the validity of the predicted labels (which were not part of the pre-established list of topics). For example, we would employ the help of human intervention to determine whether or not a song labelled as Different Interpretation would be relevant to the song in question. The results from this process could lead to further expansion and diversification of metadata, which in turn could benefit the quality of XITE’s products and channels.

Finally, we would consider including songs which are not originally written in English or which comprise more than one language in our dataset. We would start with evaluating its comprehension of the Dutch language, given that XITE is based in the Netherlands and therefore has a large database of Dutch-language music videos. At the time of writing, ChatGPT’s latest update was carried out in September 2023 - this includes support for a variety of languages except for Dutch (OpenAI (2023b)). Despite this, a conversation with ChatGPT confirmed that it is indeed able to understand and write in Dutch. However, because ChatGPT is predominantly an English language model, it is not guaranteed that it would obtain similar results with a different language (a claim which is supported by Lai et al. (2023)); an additional challenge is presented when we take lyrics into consideration, which are structured differently from the prose structure of the data on which ChatGPT has been trained. These observations could lead to an interesting study into the capabilities of ChatGPT to perform topic extraction on song lyrics on a language which is not officially configured by OpenAI.

Appendix A

Appendix

In this section we present 5 tables:

1. Overlapping songs between Knowledge and Lyrics prompts with a Perfect Match score - ChatGPT 3.5 (Closed Method)
2. Overlapping songs between Knowledge and Lyrics prompts with a Perfect Match score - ChatGPT 4 (Closed Method)
3. Overlapping songs between Knowledge and Lyrics prompts with a Perfect Match score - ChatGPT3.5 (Semi-Closed Method)
4. Overlapping songs between Knowledge and Lyrics prompts with a Perfect Match score - ChatGPT3.5 (Semi-Closed Method)

Each table shows the songs which obtained a Perfect Match score. The songs highlighted in bold are the ones that are present in *both* tables for the respective method (Closed and Semi-Closed).

The fifth table presents information regarding the 29 songs which gained a Perfect Match score for both ChatGPT versions (Closed Method). These are: 5 most frequent words, topic word count (text), topic word count (title), the total count of words and the count of unique words.

Track	Artist
All I Can Do	Dolly Parton
Crazy Little Thing Called Love	Queen
Love Song	Lana del Rey
Meaning of Life	Kelly Clarkson
Best of Me	Alicia Keys
Way Back	TLC
Bobby Jean	Bruce Springsteen
Grigio Girls	Lady Gaga
Best Friends	The Weeknd
Candy Shop	50 Cent
Lemon Song	Led Zeppelin
Birthday cake	Rihanna
Everybody Hurts	R.E.M.
Lithium	Nirvana
Betcha Gon' Know (The Prologue)	Mariah Carey
Between the Cheats	Amy Winehouse
High Infidelity	Taylor Swift
Back to Black	Amy Winehouse
Let Somebody Go	Coldplay
Cold Turkey	John Lennon
Gasoline	The Weeknd
40	U2
Act of God	Prince
Brightest Morning Star	Britney Spears
Please God Don't Tell Anyone	Jack White
Freedom	Justin Bieber
Praise God	Kanye West
Hammer To Fall	Queen
Art of Dying	George Harrison
Fade to Black	Metallica
Everybody Dies	Billie Eilish
Kill or Be Killed	Muse
911 Mr Lonely	Tyler, the Creator
Alien	Britney Spears
Eleanor Rigby	Beatles
Alone	Burna Boy
Never Felt So Alone	Billie Eilish ft Labrinth

Table A.1: Overlapping songs between Knowledge and Lyrics prompts with a Perfect Match score - ChatGPT 3.5 (Closed Method)

Track	Artist
All I Can Do	Dolly Parton
Crazy Little Thing Called Love	Queen
Love Song	Lana del Rey
Meaning of Life	Kelly Clarkson
After All	Elton John ft Charlie Puth
All Night Parking	Adele
Best of Me	Alicia Keys
Halley's Comet	Billie Eilish
Overpass Graffiti	Ed Sheeran
Maureen	Sade
Bobby Jean	Bruce Springsteen
Grigio Girls	Lady Gaga
Best Friends	The Weeknd
Candy Shop	50 Cent
Birthday cake	Rihanna
Cuff It	Beyonce
Lithium	Nirvana
My Mind & Me	Selena Gomez
Betcha Gon' Know (The Prologue)	Mariah Carey
Between the Cheats	Amy Winehouse
Confession	Destiny's Child
I Heard You're Married	The Weeknd ft. Lil Wayne
She Don't Know	Carrie Underwood
High Infidelity	Taylor Swift
Don't You	Taylor Swift
Let Somebody Go	Coldplay
Hits Different	Taylor Swift
A Baltimore Love Thing	50 Cent
Cold Turkey	John Lennon
Jimmy, Brian and Mike	Eminem
40	U2
Act of God	Prince
After Forever	Black Sabbath
Please God Don't Tell Anyone	Jack White
Freedom	Justin Bieber
Praise God	Kanye West
42	Coldplay
Hammer To Fall	Queen
Art of Dying	George Harrison
Everybody Dies	Billie Eilish
Ghosts Again	Depeche Mode
911 Mr Lonely	Tyler, the Creator
Alien	Britney Spears
Eleanor Rigby	Beatles
Alone	Burna Boy
Never Felt So Alone	Billie Eilish ft Labrinth

Table A.2: Overlapping songs between Knowledge and Lyrics prompts with a Perfect Match score - ChatGPT 4 (Closed Method)

Track	Act
All I Can Do	Dolly Parton
Crazy Little Thing Called Love	Queen
Love Song	Lana del Rey
Best of Me	Alicia Keys
Bobby Jean	Bruce Springsteen
Grigio Girls	Lady Gaga
Best Friends	The Weeknd
Candy Shop	50 Cent
Lemon Song	Led Zeppelin
Cuff It	Beyoncé

Table A.3: Overlapping songs between Knowledge and Lyrics prompts with a Perfect Match score - ChatGPT3.5 (Semi-Closed Method)

Track	Act
All I Can Do	Dolly Parton
Crazy Little Thing Called Love	Queen
Love Song	Lana Del Rey
Meaning of Life	Kelly Clarkson
After All	Elton John ft. Charlie Puth
All Night Parking	Adele
Best of Me	Alicia Keys
Halley's Comet	Billie Eilish
Maureen	Sade
Way Back	TLC
Bobby Jean	Bruce Springsteen
Grigio Girls	Lady Gaga
Best Friends	The Weeknd
Candy Shop	50 Cent
Birthday cake	Rihanna
Cuff It	Beyoncé
Summer Renaissance	Beyonce
Everybody Hurts	R.E.M.
Betcha Gon' Know (The Prologue)	Mariah Carey
Between the Cheats	Amy Winehouse
I Heard You're Married	The Weeknd ft. Lil Wayne
She Don't Know	Carrie Underwood
High Infidelity	Taylor Swift
Don't You	Taylor Swift
Let Somebody Go	Coldplay
Hits Different	Taylor Swift
Blue Banisters	Lana del Rey
A Baltimore Love Thing	50 Cent
Jimmy, Brian and Mike	Eminem
40	U2
After Forever	Black Sabbath
42	Coldplay
911 Mr Lonely	Tyler, the Creator
Alien	Britney Spears
Eleanor Rigby	Beatles
Alone	Burna Boy
Never Felt So Alone	Billie Eilish ft Labrinth

Table A.4: Overlapping songs between Knowledge and Lyrics prompts with a Perfect Match score - ChatGPT4 (Semi-Closed Method)

Track	Artist	Pre-2021	5 Most frequent words
All I Can Do	Dolly Parton	Yes	'keep', 'love', 'falling', 'youall', 'letting'
Crazy Little Thing Called Love	Queen	Yes	'thing', 'called', 'yeah', 'little', 'love'
Love Song	Lana del Rey	Yes	'i'm', 'make', 'car', 'safe', 'taste'
Meaning of Life	Kelly Clarkson	Yes	'llfe', 'meaning', 'show', 'feel', 'meaning'
Best of Me	Alicia Keys	Yes	'forever', 'get', 'gotta', 'rock', 'best'
Bobby Jean	Bruce Springsteen	Yes	'could', 'wished', 'ever', 'say', 'goodbye'
Grigio Girls	Lady Gaga	Yes	'ooh', 'pinotpinot', 'grigio', 'oh', 'make'
Best Friends	The Weeknd	Yes	'yeah', 'best', 'friend', 'oh', 'friends'
Candy Shop	50 Cent	Yes	'til', 'spot', 'take', 'candy', 'going'
Birthday Cake	Rihanna	Yes	'cake', 'name', 'put', 'cakecake', 'wanna'
Cuff It	Beyonce	No	'fuck', 'yeah', 'go', 'gon", 'night'
Lithium	Nirvana	Yes	'i'm", 'gonna', 'yeahyeah', 'cracki', 'cause'
Between the Cheats	Amy Winehouse	Yes	'ohhh', 'ooh', 'wooh', 'hoo', 'cheats'
High Infidelity	Taylor Swift	No	'dancing', 'around', 'know', 'infidelity', 'really'
Let Somebody Go	Coldplay	Yes	'oh', 'let', 'somebody', 'love', 'hurts'
Cold Turkey	John Lennon	Yes	'turkey', 'got', 'see', 'oh', 'wish'
Jesus Lord	Kanye West	No	'jesus', 'lord', 'like', 'know', 'someone'
40	U2	Yes	'sing', 'long', 'song', 'new'
Act of God	Prince	Yes	'act', 'want', 'godcall', 'except', 'god'
Praise God	Kanye West	No	'get', 'i'm', 'right', 'let's', 'still'
Hammer To Fall	Queen	Yes	'hammer', 'time', 'oh', 'one', 'know'
Art of Dying	George Harrison	Yes	'come', 'time', 'art', 'there'll', 'us'
Everybody Dies	Billie Eilish	No	'everybody', 'might', 'wanna', 'dies', 'like'
911 Mr Lonely	Tyler, the Creator*	Yes	'call', 'one', 'nine', 'never', 'lonely'
Alien	Britney Spears	No	'alone', 'not', 'like', 'stars', 'sky'
Heartbreak Hotel	Elvis Presley	Yes	'lonely', 'could', 'get', 'babywell', 'they'll'
Eleanor Rigby	Beatles	Yes	'lonely', 'look', 'ah', 'where', 'people'
Alone	Burna Boy	No	'leave', 'go', 'fit', 'make', 'body'
Never Felt So Alone	Billie Eilish ft Labrinth	No	'alone', 'felt', 'never', 'na', 'oh'

(a) Best-performing tracks from Closed Method: Track, Artist, Pre-2021, Most common words

Track	Topic word count (text)	Topic word count (title)	Total words	Unique words
All I Can Do	1	0	130	55
Crazy Little Thing Called Love	1	0	174	60
Love Song	0	1	131	72
Meaning of Life	0	0	176	78
Best of Me	0	0	89	78
Bobby Jean	0	0	114	62
Grigio Girls	0	0	144	71
Best Friends	2	1	154	74
Candy Shop	0	0	282	161
Birthday Cake	0	0	99	55
Cuff It	0	1	298	127
Lithium	0	0	154	49
Between the Cheats	1	1	160	74
High Infidelity	1	1	202	84
Let Somebody Go	0	0	129	62
Cold Turkey	0	1	61	50
Jesus Lord	1	1	740	459
40	0	0	59	26
Act of God	1	1	155	90
Praise God	0	1	347	201
Hammer To Fall	0	0	107	93
Art of Dying	0	1	66	53
Everybody Dies	1	1	75	66
911 Mr Lonely	1	1	366	236
Alien	1	0	127	47
Heartbreak Hotel	1	0	104	64
Eleanor Rigby	1	0	80	46
Alone	0	1	170	104
Never Felt So Alone	1	1	123	56

(b) Best-performing tracks from Closed Method (cont.): Track, Topic word count (text), Topic word count (title), Total word count and Unique word count

Bibliography

- a. URL <https://www.songfacts.com/facts/50-cent/a-baltimore-love-thing>.
 - b. URL <https://www.songfacts.com/facts/david-bowie/beauty-and-the-beast>.
 - c. URL <https://www.songfacts.com/facts/dangelo/brown-sugar>.
 - d. URL <https://www.songfacts.com/facts/red-hot-chili-peppers/tippa-my-tongue>.
 - e. URL <https://www.songfacts.com/facts/ed-sheeran/bad-habits>.
- R. Ali, O. Y. Tang, I. D. Connolly, P. L. Zadnik Sullivan, J. H. Shin, J. S. Fridley, W. F. Asaad, D. Cielo, A. A. Oyelese, C. E. Doberstein, et al. Performance of chatgpt and gpt-4 on neurosurgery written board examinations. *medRxiv*, pages 2023–03, 2023.
- T. Brown, B. Mann, N. Ryder, and M. Subbiah. Language models are few shot learners. *arXiv preprint arXiv:2005.14165*, 2020.
- J. Chen, P. Ying, and M. Zou. Improving music recommendation by incorporating social influence. *Multimedia Tools and Applications*, 78:2667–2687, 2019.
- K. Choi, J. H. Lee, and J. S. Downie. What is this song about anyway?: Automatic classification of subject using user interpretations and lyrics. In *IEEE/ACM Joint Conference on Digital Libraries*, pages 453–454. IEEE, 2014.
- T. Denzler. What’s in a song? using lda to find topics in over 120,000 songs, May 2021. URL <https://tim-denzler.medium.com/whats-in-a-song-using-lda-to-find-topics-in-over-120-000-songs-53785767b692>.
- B. Ding, C. Qin, L. Liu, L. Bing, S. Joty, and B. Li. Is gpt-3 a good data annotator? *arXiv preprint arXiv:2212.10450*, 2022.
- F. Kleedorfer, P. Knees, and T. Pohle. Oh oh oh whoah! towards automatic topic detection in song lyrics. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, pages 287–292, 2008.
- T. Knot. URL <https://www.theknot.com/content/taylor-swift-1989-album-wedding-songs>.
- V. D. Lai, N. T. Ngo, A. P. B. Veyseh, H. Man, F. Dernoncourt, T. Bui, and T. H. Nguyen. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*, 2023.
- X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. T. Tang. Gpt understands, too. *arXiv:2103.10385*, 2021.

- Z. Liu, X. Yu, L. Zhang, Z. Wu, C. Cao, H. Dai, L. Zhao, W. Liu, D. Shen, Q. Li, et al. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032*, 2023.
- B. Logan, A. Kositsky, and P. Moreno. Semantic analysis of song lyrics. In *2004 IEEE International Conference on Multimedia and Expo (ICME)(IEEE Cat. No. 04TH8763)*, volume 2, pages 827–830. IEEE, 2004.
- M. Modari, K. Blagec, F. Haberl, and M. Samwald. Gpt-3 models are poor few-shot learners in the biomedical domain. *arXiv preprint arXiv: 2109.02555*, 2021.
- H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.
- OpenAI. Gpt-4 technical report, 2023a.
- OpenAI. Chatgpt release notes, 2023b.
- V. Papazoglou and R. Gaizauskas. Using listeners’ interpretations in topic classification of song lyrics. In *Proceedings of the 2nd Workshop on NLP for Music and Spoken Audio (NLP4MusA)*, pages 22–26, 2021.
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- J. Ye and X. Chen. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv:2303.10420*, 2023.
- W. X. Zhao and K. Zhou. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.