Important Note. This is a pre-print that has not yet been peer-reviewed (Dated 24.07.2024).

The Use of Large Language Models for Qualitative Research: DECOTA

Lois Player [1,2,3*], Ryan Hughes [2,4], Kaloyan Mitev [1,3], Lorraine Whitmarsh [1,3], Christina Demski [1,3], Nicholas Nash [1], Trisevgeni Papakonstantinou [5], Mark Wilson [1,3]

[1] Department of Psychology, University of Bath

[2] Centre for Doctoral Training in Advanced Automotive Propulsion Systems (AAPS), University of Bath

[3] Centre for Climate Change and Social Transformations (CAST), University of Bath

[4] Department of Mechanical Engineering, University of Bath

[5] Department of Experimental Psychology, University College London

* Corresponding Author

Author Footnotes.

Acknowledgements. We would like to thank Lauren Towler from University College London for her time discussing the practicalities of Thematic Analysis.

Funding. Lois Player and Ryan Hughes are supported by a scholarship from the EPSRC Centre for Doctoral Training in Advanced Automotive Propulsion Systems (AAPS), under the project EP/S023364/1.

Authors Note. Kaloyan Mitev is now at the Joint Research Centre, European Commission. The views expressed in this publication are those of the author and do not reflect the official position of the European Commission.

Correspondence should be sent to Lois Player (lkp32@bath.ac.uk) at the Department of Psychology, University of Bath, Claverton Down, BA2 7AY.

Abstract

Machine-assisted approaches for free-text analysis are rising in popularity, owing to a growing need to rapidly analyse large volumes of qualitative data. In both research and policy settings, these approaches have promise in providing timely insights into public perceptions and enabling policymakers to understand their community's needs. However, current approaches still require expert human interpretation – posing a financial and practical barrier for those outside of academia. For the first time, we propose and validate the Deep Computational Text Analyser (DECOTA) - a novel Machine Learning methodology that automatically analyses large free-text datasets and outputs concise themes. Building on Structural Topic Modelling (STM) approaches, we used two fine-tuned Large Language Models (LLMs) and sentence transformers to automatically derive 'codes' and their corresponding 'themes', as in Inductive Thematic Analysis. To automate the process, we designed and validated a novel algorithm to choose the optimal number of 'topics' following STM. This approach automatically derives key codes and themes from free-text data, the prevalence of each code, and how prevalence varies with covariates such as age and gender. Each code is accompanied by three representative quotes. Four datasets previously analysed using Thematic Analysis were triangulated with DECOTA's codes and themes. We found that DECOTA is approximately 378 times faster and 1920 times cheaper than human coding, and consistently yields codes in agreement with or complementary to human coding (averaging 91.6% for codes, and 90% for themes). The implications for evidence-based policy development, public engagement with policymaking, and the development of psychometric measures are discussed.

Accessible Abstract

Computational approaches are increasingly being used to quickly process large volumes of free-text data. These approaches hold promise in helping academics study public perceptions, and policymakers understand their community's needs. However, current methods still require expert human interpretation, which can be costly and impractical. In this paper, we developed the Deep Computational Text Analyser (DECOTA), a novel Machine Learning tool designed to automatically analyse large free-text datasets to produce concise 'themes' within the data. DECOTA uses several custom-trained models to detect themes and sub-themes within the data, as a human may do when categorising free-text responses. Our approach gives information about how common each sub-theme and theme is, how common they are amongst different demographic groups, and offers example quotes. We compared how similar DECOTA's analysis was to human coders, using four example free-text datasets. DECOTA's outputs were highly consistent with human analysis, detecting 91.6% of all human sub-themes, and 90% of the humans' themes. We noted that DECOTA was approximately 378 times faster and 1920 times cheaper than human analysis. The potential uses of this methodology for policymakers and academics are discussed.

Keywords: Free-Text Analysis, Large Language Models, Machine Learning, Natural Language Processing, Topic Modelling

1. Introduction

Qualitative data can provide rich insights into a person's perceptions and lived experiences, in a way that quantitative data cannot (Moser & Korstjens, 2017). One form of qualitative data that is often collected as part of wider surveys is free-text data, which enables participants to give further insight into their thoughts not otherwise captured by quantitative questions. The capacity to easily collect large amounts of free-text data in these surveys is useful in evaluating public perceptions of emerging challenges, policies, and products (Rich et al., 2013), where little is known about the general population's views. This data can help researchers identify important themes and hypotheses to inform future research and allows the participant to guide the content based on their experience, unlike quantitative insights (Tenny et al., 2024). Similarly, free-text data is useful in generating hypotheses inductively, in longitudinal studies where free-text can provide detailed insights into changes over time, or as a first insight into a population's thoughts to inform quantitative work or psychometric measures.

Current approaches to analyse free-text data, such as Thematic Analysis (TA; Braun & Clarke, 2006), are typically time-consuming and require expert skills. TA is an iterative and recursive process that involves several qualitative researchers (Byrne, 2022), and derives sub-themes ('codes') and themes from a dataset. Research suggests that TA generally suits 10-50 free-text responses, with the upper limit of responses being around 400 due to issues with excessive cognitive load, time constraints, and consistency of coding (Fugard & Potts, 2015). However, there are often scenarios where response numbers exceed this, or data is overly complex and unstructured. Such limitations mean that those working outside of academic settings often lack the tools, time, and skilled personnel to analyse large volumes of free-text data, and much is left unanalysed. Certain sectors face particular challenges with this. In healthcare, for example, it is estimated that over 80% of data is unstructured (Kong, 2019), meaning free-text without a pre-defined format (Pisaneschi, 2024). In Government, vast amounts of unstructured data are collected in the form of public comments and consultations, with much of this data reportedly going unanalysed due to lack of financial resource, expertise and time (National Audit Office, 2019). If properly analysed, this data holds promise in understanding public perceptions of incoming policies and using community insights to design fair, effective and therefore well-accepted policies.

1.1 Advances in Computational Approaches

Advances in computational approaches have begun to address this lack of capacity to analyse free-text data. Natural Language Processing (NLP), for example, seeks to automate the extraction of insights from free-text data, and includes interpreting the sentiment of a text (i.e., positive, negative, neutral), drawing out key topics, and summarising text (Lee, 2023). NLP can be particularly useful in emerging fields where an inductive exploratory approach is required (Crowston et al., 2012). Topic Modelling is a common NLP methodology, which identifies patterns (‘topics’) within free-text responses (‘documents’). Broadly, it uncovers the latent themes within a large amount of text documents – enabling researchers to summarise and categorise free-text data into coherent themes (Nikolenko et al., 2017). Topic Modelling is particularly useful in identifying key themes in an exploratory way, and has been used to analyse data across several domains, including to understand perceptions of Brexit (del Gobbo et al., 2021), carbon taxes (Povitkina et al., 2021), and the COVID-19 pandemic (Wright et al., 2022). Notably, these are all instances where fast, large-scale analysis was required to inductively understand an emerging societal challenge. Of recent, there is increasing interest amongst governmental bodies in the use of Artificial Intelligence (AI) and Machine Learning (ML) to help analyse free-text data. The UK Government’s Department for Transport (DfT) recently published a report exploring the public’s perceptions of using AI to analyse consultation policy data, with the intention to reduce the amount of resources and time currently allotted to the interpretation of free-text data (Department for Transport, 2024). The report found that whilst accuracy and potential for bias were salient concerns, most people felt comfortable with this use-case – often believing that AI would analyse free-text data faster and more efficiently than a human. Within the academic sphere, NLP methodologies may also allow large amounts of free-text data to inform the development of novel psychometric measures or quantitative work, where previously only small qualitative interviews or focus groups could be used. Existing literature suggests that Topic Modelling typically captures similar findings to traditional TA approaches (Towler et al., 2022), though can be less detailed (Nelson et al., 2021).

Topic Modelling approaches are increasingly being used in social science research (Nikolenko et al., 2017; Pandur et al., 2020), and have been shown to be approximately four times faster than TA alone (Towler et al., 2022). Several types of topic models exist, for example Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA) and Structural Topic Models (STMs). STM models extend traditional LDA approaches by incorporating metadata – enabling researchers to model how the prevalence of certain topics differs across covariates (e.g., age and gender). This added insight can be particularly useful in applied contexts, where different populations may have differing perceptions or concerns.

Despite this, Topic Modelling still requires considerable interpretation of the output by a skilled qualitative researcher (Lee et al., 2017). This human intervention is required at several stages to determine:

1. The correct number of topics to split the data into, based on model fit graphs generated during the analysis.

2. The ‘codes’ (sub-themes) that underlie the data. These are derived from (a) words and phrases that have a *high probability* of occurring in a certain topic and (b) words and phrases that are *exclusive* to a certain topic.
3. The broader ‘themes’ that represent the codes.

Owing to this need for human interpretation, Topic Modelling approaches alone are often inaccessible for non-experts. For example, Towler et al. (2022) noted that STM still required approximately 40 hours of expert interpretation. This means that those outside of academia, such as in policymaking environments, must still hire external experts to conduct this analysis, creating a similar time and financial barrier as with more traditional qualitative approaches.

1.2 Large Language Models

The recent rise of Large Language Models (LLMs) could have vast implications for the interpretation of free-text data. An LLM is a type of machine learning model designed to understand, generate and interact with human language (IBM, 2024a). The models are built using transformer neural networks, are trained on vast amounts of online data, and are often capable of self-learning (AWS, 2024) – making them proficient at a wide variety of tasks from text generation, translation, summarisation, and answering questions across multiple contexts.

Alongside their general capability, LLMs can also be ‘fine-tuned’ to excel at a particular task (Thirunavukarasu et al., 2023). This involves taking the pre-trained model and further training it on a smaller, specialised dataset to highlight the types of output you would expect from a specific input (Church et al., 2021). As a result, fine-tuned models can give improved response quality and contextually appropriate outputs, with fewer irrelevant outputs, in less processing time (Ferrer, 2024). In this way, fine-tuned LLMs hold promise in conducting repetitive tasks where inputs and outputs are always structured the same, in a human-like manner.

Given this, we propose that fine-tuned LLMs can aid in interpreting the outputs of existing NLP methods, such as topic models – thereby removing the need for expert human interpretation. This methodology aims to further automate existing NLP methods, making them more accessible for use outside of the academic sphere, as well as vastly reducing the financial and time cost of analysing free-text data. It is important to note that this methodology does not seek to replace detailed, interpretative human analyses of free-text data, but rather to summarise and describe themes in scenarios where data may otherwise go unanalysed. Examples of this include when datasets are too large to analyse by hand, in scenarios where rapid analysis is necessary, or to complement more in-depth analysis methods (e.g., alongside discourse analyses of smaller interviews or focus groups).

1.3 Overview of the Proposed Method

For the first time, we propose and validate the Deep Computational Text Analyser (DECOTA) - a novel Machine Learning methodology to automatically analyse large free-text datasets and output concise codes and themes. We aimed to loosely mirror the methodology of an inductive TA, by producing initial 'codes' from the data, which are grouped into broader 'themes'. However, in contrast to TA, our proposed method seeks to summarise, not interpret, free-text data – to mitigate against potential interpretation biases in LLMs. DECOTA also gives additional insight on whether certain themes appear more amongst certain demographic variables (e.g., age and gender).

Firstly, we developed a rules-based algorithm to decipher an adequate number of topics (K) for the STM. Currently, there are no methodologies to automatically estimate an optimal number of topics, with researchers generally estimating using model fit indices generated by R's *SearchK* function. Some methodologies have sought to provide a useful starting number for the number of topics (e.g., Mimno & Lee, 2014), but not suggested a suitable topic number choice. The topic choice algorithm was validated on nine published papers, in which humans had previously chosen the number of topics. The methodology then uses STM to generate initial 'topics' and uses a fine-tuned GPT-3.5 LLM to generate coherent 'codes' from the STM output. This is followed by a cluster analysis to group similar codes together, and a final fine-tuned GPT-3.5 model to name the clusters of codes into final themes. DECOTA was validated using a formal triangulation with four independent free-text datasets, which had been analysed using an inductive TA.

2. Method

2.1 Overview of the Method

When designing DECOTA, we sought to broadly mimic an inductive Thematic Analysis (TA; Braun & Clarke, 2006) – one of the most widely-used qualitative coding approaches (Naeem et al., 2023). This involves using an exploratory approach to create initial codes (sub-themes), which are eventually grouped into broader themes. Our process followed a six-step approach:

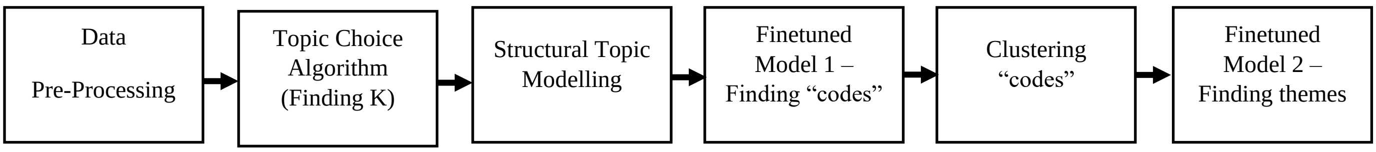
1. **Data Pre-Processing**. Ensuring the data was in the correct format to conduct the initial Structural Topic Model.
2. **Topic Choice Algorithm**. Uses a rules-based algorithm to choose an optimal number of topics (K), based on the graphs outputted from R's *SearchK* function. This new algorithm was validated with nine examples from previous literature, for which a human had determined K.

3. **Structural Topic Modelling**: Creates a Structural Topic Model (STM) using R, outputting (1) words that were exclusive to a topic, (2) list of representative quotes for each topic. This included modelling the relationship between topics and covariates (gender and age).
4. **Fine-Tuned Model 1:** Takes the STM outputs and using a fine-tuned GPT-3.5 model to create initial codes from the data.
5. **Clustering:** Uses the BERT (Bidirectional Encoder Representations from Transformers) sentence transformer to cluster similar codes together, into groups based on how similar words and phrases in the code are. A sentence transformer converts sentences into numerical representations, which capture their semantic meaning (Efimov, 2023).
6. **Fine-Tuned Model 2**: Takes the clusters of codes and uses a fine-tuned GPT 3.5 model to name them as overarching themes.

We then validated DECOTA on four independent free-text datasets on varying topics. All four datasets had been previously analysed by human coders using an inductive TA approach. Two coders performed a methodological and investigator triangulation, to compare the results from the machine and human analyses.

All code for analysis and data used to train our two fine-tuned models are openly accessible. This study received full ethical approval from the relevant University's ethics board (code 1786-1699). We used R version 4.3.2, and Python version 3.11.8 for all analyses.

Figure 1. *A flow diagram of the methodology*.



2.2 Data Pre-Processing

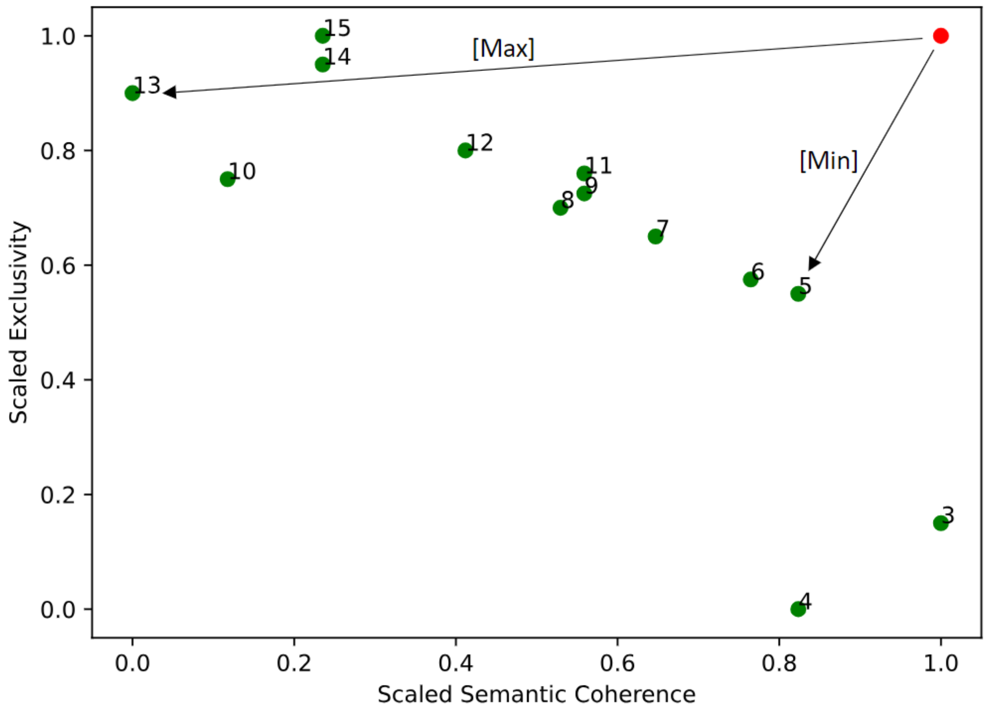
The free-text data were pre-processed using R. This involved removing missing values in the covariates, and instances where participants had not answered the question (e.g., written 'none' or 'N/A'). Covariates were converted to the correct format using base R functions. Free-text responses were split by delimiters, to capture only a singular idea per response. The delimiters chosen to split by were spaces, tabs, full stops, dashes or hyphens, semicolons, and colons. We did not split by commas, as this often did not represent the start of a new idea. The free-text data was tokenised using the *Quenteda* R package – meaning converted from a sequence of text into smaller components, and special characters removed. All free-text was stemmed using *Quenteda* (e.g., converting ‘going’ to ‘go’), and stop words removed (e.g., ‘the’ and ‘on’).

2.3 Topic Choice Algorithm

Prior to running the STM, the *SearchK* function in R (stm package) was used to create diagnostic graphs to identify the optimal number of topics (K). We tested models with 5-40 topics, to give a wide range of options for K. Previous literature suggests that the range given for K should consider the length of responses and specificity of the prompts – with shorter responses and more specific prompts indicating a smaller upper bound for K (Weston et al., 2023). Our prompts were relatively specific and response length was relatively low, hence we decided on an upper bound of 40 topics with a large range starting at 5.

The function outputs diagnostics for five key metrics: (1) *Semantic Coherence* – referring to the degree to which the top words in a given topic are meaningfully related to each other, (2) *Exclusivity* – a metric that assesses how exclusive the top words of a topic are to that topic alone, as opposed to appearing frequently across multiple topics, (3) *Held-Out Likelihood* – a measure of the generalizability of a statistical model, (4) *Lower Bound* – a measure of variance, and (5) *Residuals* – referring to the difference between observed frequencies of words and the values predicted by the model. Existing research states that when humans are determining a value, they should balance semantic coherence and exclusivity, whilst discarding topic numbers with particularly low lower bounds, or high residuals (Weston et al., 2023). This approach prioritises balancing interpretability and distinctiveness of topics, whilst accounting for other fit statistics. Based on this logic and to aid full automation of DECOTA, we created a rules-based algorithm that mimics this human decision-making. The algorithm firstly discards K values where the lower bound or residual are below the lower quartile, then finds the point where exclusivity and semantic coherence are most balanced (i.e., both at the highest point possible). This is achieved by scaling the values between zero and one, then finding the topic which minimises the Euclidian distance between it and the co-ordinate location (1,1), as shown in Figure 2. In instances where exclusivity is not available, held-out likelihood is plotted against semantic coherence – which seeks to balance the statistical fit of the model to unseen data and interpretability.

Figure 2. *An example plot of the selection of number of topics (K=5).*



2.4 Structural Topic Modelling

A Structural Topic Modelling (STM) approach was adopted, using the *stm* R package. The main outputs of this analysis are (1) a list of words that have a high probability of occurring in each topic ('Highest Prob'), (2) a list of words that are exclusive to each topic ('FREX'), and (3) a list of 10 representative quotes for each topic. Another useful output is information about the mean prevalence of each topic, which we converted to a percentage in R. A final CSV file was outputted from this stage, containing the free-text question, FREX, High Prob, topic prevalence, covariate information, and 10 representative quotes.

2.5 Fine-Tuned Model 1

The base LLM of choice was a GPT-3.5 model, an extension of the original GPT-3 developed by Open AI (Brown et al., 2020). GPT-3.5 is a transformer-based language model with a neural network of approximately 175 billion parameters (Alarcon, 2020; Vaswani et al., 2017), and was trained on a diverse corpus of text data from books, websites and other online sources. Specifically, we used the GPT-3.5 Turbo 0613, since it is the newest, most efficient version available for fine-tuning (Open AI, 2023). We opted to use the GPT API for privacy reasons, since this does not record participant information to train on future models – as the GPT text tool does (Open AI, 2024a). In contrast to the text tool, the API also allows for automated data analysis using code, which is essential for our use case.

To create the fine-tuned model, we created a synthetic training dataset consisting of 50 example inputs ('user' prompts), and their outputs ('assistant' prompts). Existing guidance from OpenAI states that fine-tuning requires a minimum of 10 examples, with around 50 being suitable for training regimented tasks like ours (Open AI, 2024b). The user prompt began with "These are quotes in answer to the question "*X*". In less than 12 words, and focusing on "(*FREX - words that are exclusive to that topic*)", what is the single core theme of these quotes?" The user prompt then detailed a string of the 10 most representative quotes for that topic, as outputted by the STM. Using the FREX words alongside the quotes encouraged more specific summarisations of the data, and reduced potential for LLM hallucinations, defined as instances where LLMs generates outputs not connected to the input (IBM, 2024b). Early versions of the model trialled using quotes alone, and performed more poorly. The assistant prompt was the code that should be outputted from the inputted FREX and representative quotes. The system prompt was set as the default 'You are a helpful assistant'.

Table 1 shows an example extract of the training data. The full data used to train fine-tuned model 1 is accessible in the Supplementary Materials (S1). The training data content was from a range of subject areas, with mixed sentiments (i.e., some positive, some negative) to ensure the model was trained to process diverse data from any field.

Table 1. *Two examples of the training data* for *fine-tuned model 1.*

System	User	Assistant
You are a helpful assistant.	These are quotes in answer to the question "**What are the negative side effects you experience from using social media platforms such as Instagram?**". In less than 12 words, and focusing on "**compar, friend, feel, celeb, liv, achiev, constant, social**", what is the single core theme of these quotes: "**Every time I scroll through Instagram, I find myself comparing my everyday life to the highlight reels of my friends. It often leaves me feeling like my life is lacking in excitement and achievement…**"	**Constant comparison with friends and celebrities diminished self-worth.**
You are a helpful assistant.	These are quotes in answer to the question "**What are the most important elements of being a good leader?**". In less than 12 words, and focusing on "**inspir, exampl, role, lead, act, achiev, pow**", what is the single core theme of these quotes:	**Leading by example to inspire and empower achievements.**

	"The most impactful leaders are those who inspire by example; they don't just tell us what to do, they show us how it's done. To lead effectively, one must embody the change they wish to see, serving as a living inspiration to their team…"

Note. Quotes have been cropped for presentation in this table. Text in **bold** indicates that it changes for each input. The 'Assistant' column represents the target output.

The code used to call fine-tuned model 1 was written in Python using JupyterLab, and utilised the GPT-3.5 API. When choosing the 'degree of randomness' in the model's output, known as 'temperature', we opted for zero to aid the repeatability of model outputs. This ensured that the model tended to choose the most probable word or token, making outputs more consistent and deterministic (Fhirfly, 2023).

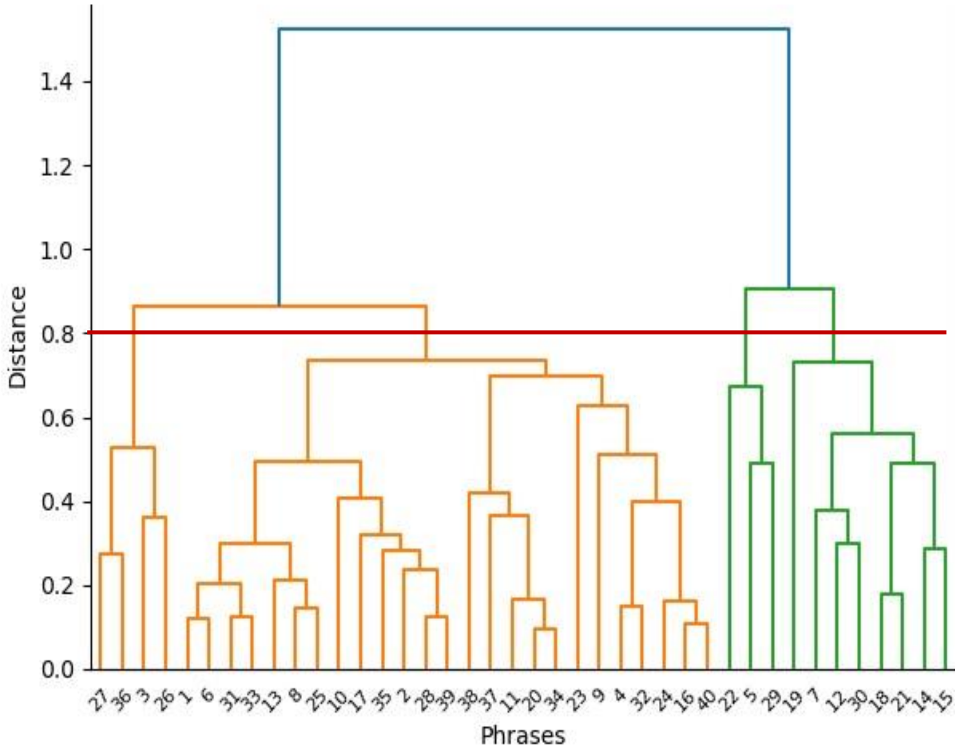
Since OpenAI do not currently allow users to directly share fine-tuned LLMs, we provide user guidance on how to use our training data to fine-tune a custom model (Supplementary Materials, S2). The required JSONL file that can be directly used for fine-tuning is also provided in the Supplementary Materials (S3).

2.6 Cluster Analysis

The cluster analysis was conducted in Python (version 3.12) and aimed to group similar codes together. The BERT sentence transformer (Devlin et al., 2019) was used to convert the text data into numerical vectors ('embeddings'), that encapsulated the semantic information of the code as a whole and represented how closely related the meaning of codes were. Cosine similarity was used to calculate the similarity scores between embeddings, and a dendrogram plotted to represent the order in which codes were grouped. Codes that were most similar were grouped first, until all texts are eventually grouped into a broader cluster. Figure 3 illustrates an example dendrogram. The branches joining represents the grouping of two codes, with the height of each join representing the distance (dissimilarity) between the clusters.

The red line in Figure 3 represents the cut-off (or threshold) value, that determines where the dendrogram is cut to form distinct clusters. Our open-access code allows the user to input their desired cut-off value, though we suggest using 0.8 as a standard since it balances number of clusters and the distinctiveness of each cluster.

Figure 3. *An example dendrogram, outputted by the cluster analysis*.



2.7 Fine-Tuned Model 2

Fine-tuned model 2 also used a GPT-3.5 Turbo 0613 model. To create the fine-tuned model, we created a synthetic training dataset consisting of 50 example inputs ('user' prompts), and their outputs ('assistant' prompts). The user prompt began with "This is a list of sub-themes that have an underlying theme, in answer to the question *'X'*. Name the underlying theme, in 6 words or less. If there are any sub-themes that are anomalies, ignore it and name the theme in line with the other sub-themes." The user prompt then detailed a list of codes, clustered together by the clustering analysis. The assistant prompt was the final theme, named by the model. The code used to call fine-tuned model 2 was written in Python, in the same environment and JupyterLab notebook as fine-tuned model 1.

Table 2 shows an example of this training data. The full data used to train fine-tuned model 2 is accessible in the Supplementary Materials (S1), alongside the formatted JSONL file to fine-tune the LLM (S3). Another BERT sentence transformer was used to select the three most representative quotes from the ten representative quotes outputted by the STM. The temperature was set to zero for fine-tuned model 2 for maximum repeatability of outputs, as in fine-tuned model 1.

Table 2. *Two examples of the training data* for *fine-tuned model 2.*

System	User	Assistant
You are a helpful assistant.	This is a list of sub-themes that have an underlying theme, in answer to the question **'What are the negative side effects you experience from using social media platforms such as Instagram?'**. Name the underlying theme, in 6 words or less. If there are any sub-themes that are anomalies, ignore it and name the theme in line with the other sub-themes. **Inadequacy from social comparison with friends & celebrities. Anxiety about missing out on experiences or achievements seen online. Concerns of being less liked due to fewer social media interactions…**	**Negative impacts of social comparison.**
You are a helpful assistant.	This is a list of sub-themes that have an underlying theme, in answer to the question **'What are the most important elements of being a good leader?'**. Name the underlying theme, in 6 words or less. If there are any sub-themes that are anomalies, ignore it and name the theme in line with the other sub-themes. **Prioritising active listening and understanding team perspectives. Effective communication both verbally and in writing. Being open to feedback from employees. Effectively negotiating with team member…**	**Role modelling good management.**

Note. Lists of codes have been cropped for presentation in this table. Text in **bold** indicates that it changes for each input. The 'Assistant' column represents the target output.

2.9 Validation Process

2.9.1 Validation of K Algorithm

DECOTA's K algorithm was compared to nine examples from previous literature. Currently, an optimal number of topics for STM is chosen using the output graphs of the *SearchK* function. We sought to validate if, when given the graphs outputted by the *SearchK* function in R, our rules-based algorithm would reliably choose the same number of topics (K) as authors did in previous research. Nine examples were found, where the authors had reported the output of their *SearchK* results (Chung et al., 2022; Kim et al., 2020; Lindstedt, 2019; Mickelsson et al., 2022; Mostafa, 2023; Pandur et al., 2020; Towler et al., 2022; Ünver & Kurnaz, 2022; Xiang, 2022). Whilst we acknowledge that in some contexts the choice of topic number is context dependent (e.g., the author may wish to prioritise some model fit indices over others; Meaney et al., 2023), we sought to show that our algorithm can consistently choose a logical and useful number of topics.

2.9.2 Validation Datasets

To validate DECOTA, we obtained four secondary independent datasets, which differed in content, number of participants, and slightly in length of each free-text response. All were secondary datasets collected and analysed for other research projects, by four different researchers. Descriptions of the datasets are highlighted in Table 3, and the raw data are available for all in the Supplementary Materials (S4).

Table 3. *Characteristics of the four datasets.*

Dataset No	Dataset Name	Free-text Question	Number of Participants	Average response length (words)
1	Drought Behaviours	Why do you try and save water at home?	1034	9.3
2	Germ Defence	What was helpful about the Germ Defence website?	1472	12.1
3	Cornwall Council	What in your view are the most important actions Cornwall Council should take to help tackle climate change?	859	12.4
4	Sustainability Actions	There is a range of actions you can take that are good for the environment or that can help lower your impact. Please name three individual actions.	1100	11.2

All datasets had been previously analysed using an inductive TA approach, producing a list of codes and themes, but had been analysed to differing levels of details – reflective of the variability in human coding. Two of the four datasets had information about the prevalence of certain codes (Datasets 1 and 3). Table 4 illustrates the basic data structure DECOTA uses. Across the sample datasets, we used gender as 1 = Male, 0 = Any other gender, and age as an exact value, since STM covariates must be either continuous or binary. Other covariates could be coded in, but for demonstrative purposes we used age and gender.

Table 4. *The format of data inputted into DECOTA.*

ID	Question_1	age	gender
1	Example free-text data	44	1

2.9.3 *Triangulation Approach*

A formal methodological and investigator triangulation was conducted, to compare the similarity of human and DECOTA coding. This approach is appropriate when results are from two different analyses, conducted by multiple different researchers (Carter et al., 2014). Two researchers independently completed eight convergence coding matrices (a codes matrix and themes matrix for each dataset). A triangulation guidance document was created and is shared in the Supplementary Materials (S5).

The triangulation involved comparing the codes and themes from DECOTA and the human analyses, and categorising them into: (1) In Agreement – conceptual convergence between codes (including a direct example), (2) Complementary – shared meaning of essence between codes (including a less linked example), (3) Dissonance – disagreement between the codes, and (4) Silence – a code is present in only one of the analyses. Another triangulation was conducted, comparing the human and DECOTA themes.

After independently conducting the triangulation, the two researchers met and discussed differences in interpretation, and resolved them to create a final version (see Supplementary Materials – S6).

2.10 *Transparency and Openness*

All data and code associated with this manuscript can be found on the Open Science Framework (OSF). To make DECOTA as accessible as possible for users, the OSF repository includes a Google Colab link, which combines all R and Python code into one script, does not require the download of any software, and contains simple instructions to use the tool. The original R and Python scripts have also been uploaded. To more closely replicate our results, these scripts should be used, as the Google Colab environment uses slightly different versions of both R and Python and therefore may yield slightly different results. Note that differences in one's R environment can also create slight variation in the number of topics deemed suitable by the *SearchK* function. To mirror our results, users can manually input our chosen topic number.

Because the validation datasets were not collected for this specific project, the covariates were not always in the required data type. For example, Datasets 2 and 4 used various age categories (categorical variables). In our analysis, we used the original data (with some categorical variables), but also provide a processed version of the data where age categories are converted into numeric by taking the mid-point of each, which will run without edits to the code (all data available in S4, and on the OSF).

3. Results

3.1 K Validation

In six of the nine examples, our algorithm chose the exact same value for K as the authors (Lindstedt, 2019; Mickelsson et al., 2022; Mostafa, 2023; Pandur et al., 2020; Towler et al., 2022; Ünver & Kurnaz, 2022). In the remaining three examples, the algorithm chose a similar value for K – with the author's choice being the algorithm's next choice (Chung et al., 2022; Kim et al., 2020; Xiang, 2022). These choices were all sensible choices for K (i.e., balancing model fit statistics, whilst prioritising semantic coherence and exclusivity), making them in broad agreement with the author's choices. Table 5 summarises the comparisons between previous authors' and the algorithm's choices of K. As illustrated in Table 5, some authors choose a range of reasonable values for K, since there is no one correct answer – further reiterating that our 'next best choice' examples were reasonable choices.

Table 5. *A comparison of previous authors' and the algorithm's topic number choices.*

Authors	Author's Choice of K	Algorithm's Choice of K	Interpretation
Lindstedt (2019)	20-25	20	Exact match
Mickelsson et al. (2022)	80	80	Exact match
Mostafa (2023)	15-20	20	Exact match
Pandur et al. (2020)	25	25	Exact match
Towler et al. (2022)	25	25	Exact match
Ünver & Kurnaz (2022)	50	50	Exact match
Chung et al. (2022)	18	12	Next best choice
Kim et al. (2020)	14	11	Next best choice
Xiang (2022)	6	8	Next best choice

3.2 DECOTA Output

For each dataset, DECOTA outputs several pieces of information. These include (1) model fit information for choosing number of topics (K) for the STM, (2) a dendrogram from the clustering analysis, (3) a final CSV file summarising all codes, themes, covariate information and representative quotes for each code.

3.2.1 Model Fit Information for Choosing K

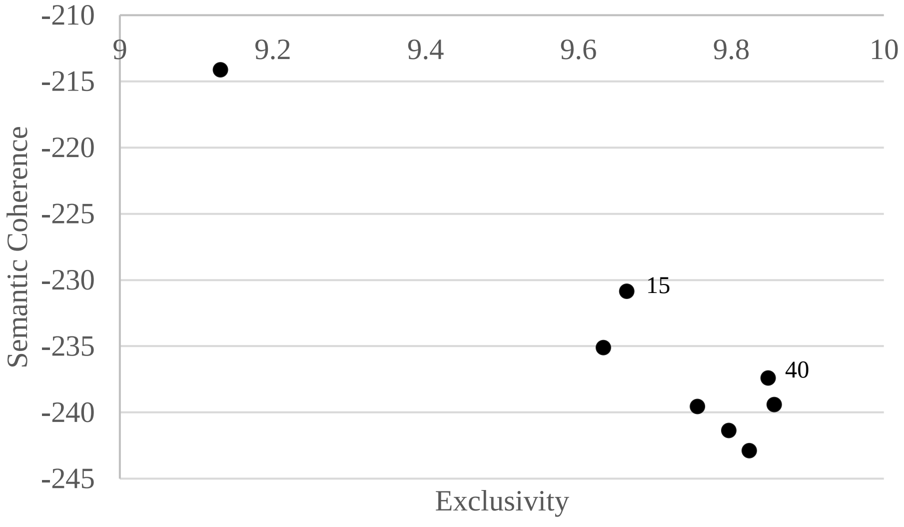
DECOTA's Topic Choice Algorithm automatically chooses a value for K that balances semantic coherence and exclusivity metrics, whilst excluding poor values for residuals, lower bound, and held-out likelihood. To make this decision, DECOTA outputs a table with all model fit indices for a range of values for K (5-40 topics). This ‘K Metrics’ table for all four datasets can be found in the Supplementary Materials (S7). For transparency, DECOTA also outputs the graphs for all model fit indices, outputted by the *SearchK* function (see Supplementary Materials – S8), and a graph plotting semantic coherence against exclusivity for 5-40 topics (see Supplementary Materials – S9). If desired, users can then check and edit the chosen value for K. For all four datasets, the algorithm chose the value for K. Table 6 shows the number of topics chosen for each dataset.

Table 6. *The algorithm’s chosen K values.*

Dataset No	Dataset Name	Chosen K Value
1	Drought Behaviours	40
2	Germ Defence	15
3	Cornwall Council	35
4	Sustainability Actions	25

The algorithm’s choices of K yielded reasonable and distinct codes for all four datasets. However, Dataset 1 may have benefitted from the choice of slightly fewer topics. Observing the exclusivity and semantic coherence plot for Dataset 1 (see Figure 4), the next best choice in balancing semantic coherence and exclusivity was 15 topics. Considering the specificity of the prompt and heterogeneity of responses for Dataset 1, 15 topics may have yielded slightly more distinct codes. Despite this, fewer topics would have meant failing to detect codes that fewer participants spoke about. Whilst DECOTA’s Topic Choice Algorithm is capable of choosing a suitable value for K, this situation illustrates that researchers may want to override the algorithm’s choice, depending on their use case.

Figure 4. *A plot of exclusivity and semantic coherence for Dataset 1.*



3.2.2 The Dendrogram

For each dataset, DECOTA outputs a dendrogram to illustrate the clustering approach outlined in section 2.6. We used a cut-off of 0.8 for three of the four datasets (Datasets 1, 2, and 3), and a cut-off of 0.6 for Dataset 4, to demonstrate that whilst 0.8 is often a good 'standard' value, free-text responses with less specific prompts and more varied responses might benefit from a lower cut-off value (i.e., a larger amount of more specific clusters). The dendrograms for all four datasets can be found in the Supplementary Materials (S10).

3.2.1 Final CSV File

DECOTA outputs a final CSV file, summarising the information from the entire six-step process. This includes the (1) overall theme name and corresponding number, (2) topic number from the original STM, (3) the code name, (4) data on the prevalence of each code in percentage, (5) covariate information on age and gender (R estimate and *p* value), and (6) the three most representative quotes for each code. Table 7 is an example output for Dataset 2 (Germ Defence), where participants were asked "What was helpful about the Germ Defence website?". Note that participants tended to mention both helpful *characteristics* of the website, alongside what *information* on the website was most helpful to them – which is reflected in our results. Figure 5 shows the breakdown from overall themes to individual codes, illustrating the prevalence of codes. The output CSV and illustrative diagrams for the other three datasets can also be found in the Supplementary Materials (S11; S12).

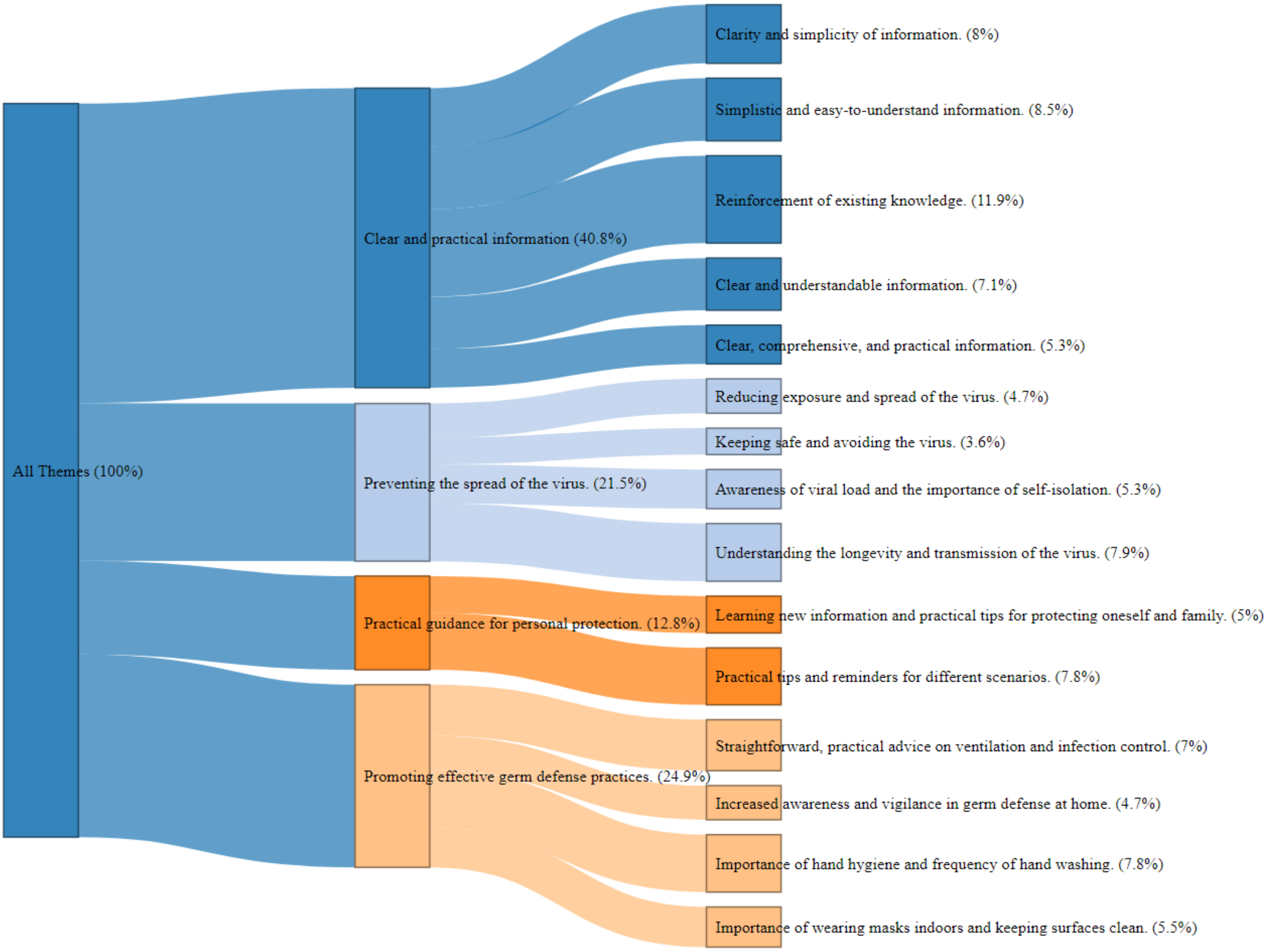
Table 7. *An example output of DECOTA, for Dataset 2 (Germ Defence)*

Theme No	Theme Name	Topic No	Code Name	Topic Prev (%)	Age (R)	Age (*p*)	Gender (R)	Gender (*p*)	Quote 1	Quote 2	Quote 3
		1	Clarity and simplicity of information.	7.98	-0.023	0.428	-0.007	0.381	The information was described in clear ways with no unnecessary jargon and without being too overwhelmingly wordy	It was a nice balance of text and pictures	The hygiene regimes and the instructions were clear
		2	Simplistic and easy-to-understand information.	8.54	-0.049	0.042	0.040	0.000	It inspired confidence that the information is a product of academic thought and based in research	Simple language and offered simple solutions	Based on research carried out at a reputable university in UK, which is reassuring
1	Clear and practical information	3	Reinforcement of existing knowledge.	11.92	0.024	0.386	0.011	0.203	It mostly confirmed what we knew already	I thought I knew it all but it made me re-evaluate my knowledge and I got a bit relaxed over my germ control as my quiz answers showed so I will step it up again after reading this	What I found helpful was when I had a chance to choose which situation I belonged to and the information that was given on that
		9	Clear and understandable information.	7.11	0.017	0.420	0.015	0.015	It made things clearer, sadly, information from media, government is not communicated well or efficiently	Digestible info, ability to share easily on social media, I learned things even as a Chemistry graduate, I didn't know	And fairly straightforward to navigate. Succinct and graphic enough
		15	Clear, comprehensive, and practical information.	5.29	-0.010	0.310	0.005	0.024	Clear and informative	Real, logical, and practical information	Breaks down information and case studies are a great way of making the information relatable

2	Preventing the spread of the virus.	6	Reducing exposure and spread of the virus.	4.69	0.000	1.000	-0.008	0.364	Helpful tips in preventing spread of virus and control exposure	And measures to reduce the amount of virus someone might get	I hadn't taken on board before that it is really beneficial to reduce the AMOUNT of Covid virus
		7	Keeping safe and avoiding the virus.	3.61	-0.023	0.421	-0.021	0.000	i.e. self isolate, avoid public transport, keep away from those who could pass viruses to me	Help to keep others safe too	If I was in an environment where I may come into contact with people who may have the virus or may contract the virus I would find the information very helpful
		12	Awareness of viral load and the importance of self-isolation.	5.33	0.016	0.607	-0.007	0.415	Emphasising that viral load increases the chance of severe illness	It reinforced my belief that the higher intensity of the viruses I am in contact with, the more chance I have of contracting the disease and the worse my illness might be	If not poss to isolate in a separate room wear a mask and stay 2 2mtrs apart from other household members and eat food in a separate room
		14	Understanding the longevity and transmission of the virus.	7.86	-0.018	0.527	-0.013	0.160	It informed me about some aspects that I have not seen or heard elsewhere, such as the length of time the virus can remain alive on surfaces	Knowing the virus can stay in the air for a long time after visitors leave	It clarified some aspects of prevention especially about the length of time the virus can live on surfaces
3	Practical guidance for personal protection.	5	Learning new information and practical tips for protecting oneself and family.	5.02	-0.065	0.063	-0.005	0.401	Although I thought I was protecting myself and my family I feel that the website has taught me to go a little bit further with what we are doing	Watching videos of real people and how they manage	Providing questionnaire for people's own situation

		13	Practical tips and reminders for different scenarios.	7.75	0.027	0.168	0.006	0.380	Good reminders of best practice	I could see it’s offering practical tips	The info on the website was good to remind me of what I should be doing and different scenarios that could happen
4	Promoting effective germ defence practices.	4	Straightforward, practical advice on ventilation and infection control.	6.95	0.034	0.095	0.019	0.002	It is a straightforward way to deliver advice on principles of good practise for infection control	General advice on ventilation, this is the problem with a lot of newer houses, nobody opens windows anymore	It gave me additional advice on how to prevent being infected
		8	Increased awareness and vigilance in germ defence at home.	4.65	0.013	0.490	-0.013	0.004	I thought I was doing pretty well with my levels of germ defence in the home but it made me realise there were areas I could do still do better with	It made me realise that I should continue with a high level of vigilance	The information was very clear and it was useful to have more information about keeping surfaces and items coming into the home germ free than I generally have found elsewhere
		10	Importance of hand hygiene and frequency of hand washing.	7.82	0.025	0.383	-0.006	0.468	Makes you think about how often you do wash hands	It made me focus on the importance of hand washing often	As time goes on I have not been washing my hands as often as I did during the first lick down
		11	Importance of wearing masks indoors and keeping surfaces clean.	5.47	0.032	0.147	-0.018	0.012	I hadn't considered wearing a mask indoors	The wearing of mask in doors the in for about glasses	I use sanitizer everywhere which I have one clip on my coat which I use if I go out anywhere plus I use face masks everywhere I go

Figure 5. *The breakdown of themes to codes, including prevalence information.*



3.3 Triangulation Results

3.3.1 Code Triangulation

The following triangulation results present first (1) how well DECOTA captured the codes humans created (see Table 8), and then (2) how well the humans captured the codes DECOTA created (see Table 9). It was important to present the comparison in both ways, since in some instances DECOTA detected codes that the humans did not. This is likely because human coding is subjective, and variable in quality and detail (Zade et al., 2018) – meaning it can also miss nuances in the data. It is important to note that for these reasons, human coding should not be viewed as the gold standard of analysis – and the aim of the following comparisons is to show that both analytical methods give the same general impression of the data.

Tables 8 and 9 illustrate these comparisons. In general, there were high levels of agreement between human and DECOTA coding. As shown in Table 8, DECOTA detected on average 91.6% of all human codes to an "agreement" or "complementary" level, across the four datasets. On average, only 8.4% of all human codes were not detected by DECOTA – often due to the human coding being more detailed and specific in some datasets (e.g., Dataset 3), or human codes representing only a very small percentage of the data. For example, the codes missed by DECOTA in Dataset 1 represented less than 0.1% of responses per code. Since DECOTA begins by using STM, this approach detects the most prevalent 'topics' in the data, and therefore will often not detect these particularly uncommon codes.

As shown in Table 9, human coders detected on average 96.3% of all codes that DECOTA did. Across the four datasets, there were 3.7% of DECOTA codes that humans did not detect, showing that there are instances where DECOTA captured elements of the data that the humans did not. There were no 'dissonant' codes in any datasets, for either comparison.

Table 8. *Comparison of human codes, as detected by DECOTA.*

Dataset No.	Dataset	**Agreement + Complementary (%)**	Agreement (%)	Complementary (%)	Silent – DECOTA Missed (%)
1	Drought Behaviours	**85.7**	64.3	21.4	14.3
2	Germ Defence	**100.0**	80.0	20.0	0.0
3	Cornwall Council	**91.6**	83.3	8.3	8.3
4	Sustainability Actions	**88.9**	71.1	17.8	11.1
	Average	**91.6**	74.68	16.9	8.4

Table 9. *Comparison of DECOTA codes, detected by humans.*

Dataset No.	Dataset	**Agreement + Complementary (%)**	Agreement (%)	Complementary (%)	Silent – Humans Missed (%)
1	Drought Behaviours	**95.0**	82.5	12.5	5.0
2	Germ Defence	**100.0**	80.0	20.0	0.0
3	Cornwall Council	**94.3**	82.9	11.4	5.7
4	Sustainability Actions	**96.0**	80.0	16.0	4.0
	Average	**96.3**	81.4	15.0	3.7

3.4.2 Theme Triangulation

A formal triangulation was also conducted for the broad themes outputted by DECOTA. In general, there was high agreement between the human and DECOTA themes. A shown in Table 10, DECOTA captured 90% of human codes to an 'agreement' or 'complementary' level, with 10% silent themes. Similarly, the humans captured an average of 89.9% of DECOTA's themes, with 12.4% of DECOTA's themes not appearing in the human analysis (see Table 11). There were no 'dissonant' themes in any datasets, for either comparison.

As expected, these values illustrate more differences between the human and machine themes than there were between codes, likely because there are many diverse ways that codes can be grouped together into broad themes. Notably, in datasets with more specific prompts and less variability in responses (e.g., Datasets 1 and 2), the theming was more similar – since there are fewer potential ways to sensibly group the codes. Datasets 3 and 4 both had very broad prompts with more variability in responses, making the DECOTA / human coding less similar. For example, in Dataset 4, DECOTA had a theme for "sustainable living". This theme could encompass many environmental actions and was therefore not easily matched to any of the slightly more specific human themes. Depending on the user's use case, DECOTA could be amended to use a lower cut-off score (see section 2.6), to yield more specific themes, if the user desired the output to have a higher number of more specific codes.

Table 10. *Comparison of human themes, as detected by DECOTA.*

Dataset No.	Dataset	**Agreement + Complementary (%)**	Agreement (%)	Complementary (%)	Silent – DECOTA Missed (%)
1	Drought Behaviours	**100.0**	100.0	0.0	0.0
2	Germ Defence	**100.0**	66.6	33.3	0.0
3	Cornwall Council	**80.0**	70.0	10.0	20.0
4	Sustainability Actions	**80.0**	70.0	10.0	20.0
	Average	**90.0**	76.7	13.3	10.0

Table 11. *Comparison of DECOTA themes, detected by humans.*

Dataset No.	Dataset	**Agreement + Complementary (%)**	Agreement (%)	Complementary (%)	Silent – Humans Missed (%)
1	Drought Behaviours	**100.0**	100.0	0.0	0.0
2	Germ Defence	**100.0**	75.0	25.0	0.0
3	Cornwall Council	**77.7**	66.6	11.1	22.2
4	Sustainability Actions	**81.8**	72.7	9.1	27.3
	Average	**89.9**	78.6	11.3	12.4

3.4.3 Interrater Reliability

Two researchers triangulated the human and DECOTA codes. To understand how similar their matching of DECOTA to human codes was, we calculated a simple percentage score and Cohen's Kappa, which measures the agreement between the two coders beyond what would be expected by chance alone (McHugh, 2012). This analysis shows how often the two coders *broadly* agreed on which DECOTA codes represented which human codes. We did not differentiate between differences in coding between 'in agreement' and 'complementary', as these two categories are conceptually similar, and both represent a general agreement.

Table 12 shows the results from this analysis. Generally, there was 'substantial' to 'excellent' agreement between the coders, beyond what would be expected by chance. Notably, the two datasets with the more specific prompts and less variability in responses (Datasets 1 and 2) had particularly high Kappa values above 0.8. The datasets with broader prompts and more response variability (Datasets 3 and 4) had Kappa values indicating 'substantial' agreement but were slightly lower (McHugh, 2012).

Key places where disagreements between coders occurred were in instances where either the human or DECOTA code was less specific – leaving room for differences in interpretation. For example, in Dataset 3's code "climate change action and support for local initiatives", the two coders assigned different codes as examples of local initiatives. Relatedly, disagreements also occurred when there were multiple codes for the same idea. Due to our use of STM, this happened when a certain topic was particularly prevalent in the data (e.g., "transport" in Dataset 3), and resulted in there being many correct, yet different DECOTA codes to match the human code to.

Table 12. Interrater reliability statistic for the codes.

Dataset No.	Dataset	Broad Agreement (%)	Kappa	Z	*p* Value
1	Drought Behaviours	97.68	0.861	20.40	.000
2	Germ Defence	97.33	0.912	7.93	.000
3	Cornwall Council	98.28	0.784	40.00	.000
4	Sustainability Actions	97.33	0.762	25.60	.000

Note. 'Broad agreement' includes 'in agreement' and 'complementary' codes.

Interrater reliability was also calculated for the themes, as shown in Table 13. Kappa values for the themes were understandably slightly lower than the codes, since the themes were less specific than the codes, and therefore there were more instances where multiple codes could be matched together or interpreted differently. Overall, the Kappa values ranged from substantial (Datasets 2, 3 and 4) to excellent agreement between coders (Dataset 1) (McHugh, 2012).

Table 13. Interrater reliability statistic for the themes.

Dataset No.	Dataset	Broad Agreement (%)	Kappa	Z	*p* Value
1	Drought Behaviours	100.00	1.000	4.00	.000
2	Germ Defence	83.33	0.667	2.45	.014
3	Cornwall Council	95.96	0.778	7.79	.000
4	Sustainability Actions	94.55	0.670	7.03	.000

Note. ‘Broad agreement’ includes ‘in agreement’ and ‘complementary’ themes.

3.5 Time and Cost Comparison

We calculated the mean time and cost savings of using DECOTA compared to traditional inductive thematic analysis approaches. On average across all four datasets, DECOTA took 10 minutes to run, compared to an average of 63 hours for the thematic analyses - making DECOTA approximately 378 times faster than traditional approaches. Assuming a standard Research Assistant pay of $25p/h, 63 hours of analysis would cost approximately $1575. The GPT-3.5 Turbo 0613 API is charged at ~$0.00003 per “token” for inputs, which broadly represents the number of words plus punctuation, spaces, and special characters. Based on the dataset characteristics in Table 3, we estimated that the average amount of tokens in our four datasets was approximately 13,000 for the inputs, costing $0.39 per model for each dataset. The outputs of the model ranged from having 4-9 themes, with each being approximately four tokens, and 15-40 codes, where codes tended to be around 10 tokens. If we consider an average of 6 themes (4 tokens per theme), and 30 codes (10 tokens per code), the model’s outputs would be 324 tokens. At the pricing of $0.00006 per token for outputs, the model’s outputs would cost ~$0.02 per dataset, and $0.41 overall.

We used two fine-tuned models, making the overall cost for use of the API approximately $0.82. This makes DECOTA approximately 1920 times cheaper than human thematic analysis, excluding hardware costs (e.g., computer use).

Towler et al. (2022) calculated that an STM approach took approximately 40 hours of analysis, owing to the need for human interpretation of STM outputs. This figure allows us to estimate that DECOTA is approximately 240 times faster than the current most novel computational methods. Using the same cost assumptions, DECOTA is approximately 1220 times cheaper than STM methods, and requires no expert human interpretation.

4. Discussion

For the first time, this study aimed to design and systematically validate a novel open-access Machine Learning methodology to automatically analyse large volumes of free-text data. The final tool, the Deep Computational Text Analyser (DECOTA) followed a six-step approach – including an algorithm to automatically suggest an optimal number of topics, STM, a clustering approach, and two fine-tuned LLMs. DECOTA broadly mimics an inductive TA – in its progression from raw free-text data, to codes, to broader themes. DECOTA outputs the themes and sub-themes underlying the data, alongside their prevalence, relationship to covariates, and three representative quotes for each sub-theme.

We validated DECOTA on four independent and varied free-text datasets, which had previously been analysed by researchers using an inductive TA approach. A formal triangulation showed that DECOTA was extremely proficient at capturing similar codes to the human researchers – finding an average of 91.6% of all human codes, and 90% of all human themes. When comparing, it is important to remember that human TA does not always detect all themes within a dataset, and therefore should not be considered the gold standard against which to compare. Our analyses reiterated this, showing that human coders also missed some nuance – detecting on average 96.3% of DECOTA's codes. There was variability in this between datasets, owing to some human coding being more detailed than others. This highlights the challenge of variability in rigour of human coding. Human coding also differs between analysts and even within the same analyst over time, whilst DECOTA performs more consistently between trials and over time.

Whilst it is important to validate DECOTA against current human methods, it is key to note that DECOTA's purpose is to summarise free-text data, rather than to interpret as humans often do. This is by design, to increase transparency and best represent the data – and avoid potentially biased interpretations of the data. In some instances, this led to DECOTA 'missing' themes, where the human had interpreted rather than summarised data. In Dataset 2, for example, the human theme "reducing all or nothing thinking" was not captured by DECOTA, since it required a level of interpretation. This use case differs from types of human analyses that aim to interpret tone, emotion and meaning from qualitative data, such as discourse analyses (Baker, 2023).

Whilst DECOTA yielded similar results to humans for all four datasets, the tool was marginally more effective on datasets with more focused prompts, and hence slightly less variability in responses (Datasets 1 and 2). This is likely because specific questions tend to produce text with clearer and more distinctive topics, which are more straightforward for STMs to model (Weston et al., 2023) – hence producing more nuanced 'codes'. For more general questions, the diversity of topics and potential overlap might make it challenging for STMs to capture the same level of detail. For use cases where detail is paramount, users may opt to use more specific prompts. Another consideration for collecting free-text data is prompting users to separate ideas by bullet points, full stops or in separate response boxes. In our analyses, the STM at times struggled to create separate, coherent topics when users had mentioned too many distinct ideas within one response, creating a knock-on effect to creating incoherent codes. To remedy this, we attempted to separate the data into distinct ideas by splitting by delimiters. Though mostly effective, this did not separate ideas in instances where no punctuation was used, and more distinct codes may have been created if users were instructed to isolate ideas in a specific way.

4.1 Strengths and Implications

In developing this tool, we aimed to make the analysis of large volumes of free-text data accessible to all, regardless of expertise, cost, and access to resources. In applied policy settings, DECOTA could be used to efficiently analyse free-text consultation data, giving valuable insights into the barriers community members would likely face in accepting a proposed policy. This tool would be particularly useful in situations that require the rapid implementation of policies, as an understanding of public sentiment could be analysed within minutes of being collected. For example, DECOTA may be useful in understanding public perceptions of incoming controversial climate policies such as Low Traffic Neighbourhoods (LTNs). LTNs are a transport policy in which motor vehicle traffic in residential streets is greatly reduced (Manifesto Club, 2023). Many were introduced swiftly and with little public consultation during the COVID-19 pandemic, resulting in huge public backlash (Wall, 2020), with many LTNs being later withdrawn. In this scenario, DECOTA may have provided a quick and cost-effective way to highlight the potential issues to policymakers ahead of time. In their recent report, the UK Government’s Department for Transport (DfT; 2024) found high public support for using AI to analyse consultation data. The report noted desire for human oversight during the analysis, and a comparative analysis against human analysis to assess quality – both of which DECOTA delivers. In an ideal policy-planning scenario, DECOTA may also be used alongside other mechanisms, for example smaller deliberative processes, to understand a community’s needs (Bua & Escobar, 2018). By fostering a policymaking approach that broadly considers a large amount of people’s opinions (i.e., via consultation data and DECOTA), and a more focused group of people’s more deeply (via deliberative processes), a holistic and thorough understanding of the community’s needs is developed.

Outside of the policymaking sphere, DECOTA may be useful during the development of quantitative psychometric measures. Currently, researchers rely on small focus groups, interviews, or Delphi studies to inform the types of items to include in a new measure (Brown, 2018). DECOTA could aid this process, by allowing a larger number of opinions to be captured in free-text – perhaps speeding up and reducing the resources required to develop novel questionnaires. Within private companies or in the user experience domain, our tool could usefully summarise large volumes of free-text employee or customer feedback. This use case is growing in demand, with some newly-formed startups funded with this goal in mind (Caplena, 2024), and large companies such as Amazon beginning to create AI-generated summaries of customer reviews (Amazon, 2023). Within academia, DECOTA unlocks potential for the automatic analysis of free-text data that may otherwise go unanalysed. For example, DECOTA may be a useful resource to analyse unanalysed free-text responses collected alongside quantitative questionnaires, that provide invaluable additional qualitative insights. Alternatively, devolving descriptive thematic coding to DECOTA could free up the researcher to perform more interpretive analysis, such as discourse analysis.

DECOTA's ability to understand the *prevalence* of different codes within the data is a particular strength. In policy applications, this enables policymakers to understand the issues or concerns most pertinent to their community, and therefore focus their resources when introducing a policy. For example, if a large amount of the population is concerned about a lack of appropriate public transport when introducing a road tax, policymakers may facilitate acceptance of the policy by introducing better local bus routes ahead of implementation. Another core strength of DECOTA is its ability to automatically detect which codes and themes were most prevalent amongst different demographic groups. In this paper, we illustrated this using gender and age, but any continuous or binary variables could be easily added. This feature allows users to understand if specific demographic groups have different opinions or concerns to others and could inform which groups may need further consultation or information before a policy is introduced.

4.2. Limitations and Areas for Future Work

DECOTA is the first transparent and open-access tool attempting to automatically analyse free-text data, by leveraging fine-tuned LLMs. Though shown to be effective for our validation datasets, there are several areas the tool could be further developed. Considering its potentially huge implications for policymakers, a key area of future research surrounds further developing the tool into both a fully accessible Python package, and a code-less website interface for widespread use. Future work may also use our open-access training data to trial our methodology on other fully open-access LLMs, as the use of our GPT model currently requires an Open AI subscription. On a more granular level, our two fine-tuned LLMs were trained using domain-general synthetic data – making the tool proficient at creating codes and themes from any topic area. If users desired the tool to be more familiar with a particular topic area, for use with free-text data from a specific specialist subject, it may be preferable to create further training data from the desired domain area to further train the LLMs with expert knowledge and language. Equally, we trained our fine-tuned LLMs on English examples, so it is unclear if DECOTA would be as effective on free-text data in other languages. Future research could test the effectiveness of DECOTA with domain-specific training data, and training data in other languages.

A challenge with DECOTA initially relying on STM is in its tendency to create new ‘topics’, if two topics are often mentioned together. For example, in Dataset 3, the idea of ‘transport’ was seemingly mentioned often both alone, and in tandem with other prevalent ideas such as ‘waste management’. Since STMs cluster words and phrases into groups based partly on them frequently occurring together (Mu et al., 2022) – this created many codes which mention ‘transport’ in conjunction with other prevalent codes. Whilst this may be somewhat useful in communicating which topics were often mentioned together, it does create some unnecessary repetition of codes. This challenge may be mitigated in future work by advising participants in free-text responses to input one idea per response box or use bullet points to separate ideas – making it easier to split by idea in the analysis. A further challenge DECOTA faces in using STM is that topic models, by design, do not seek out topics particularly low in prevalence. This can result in very uncommon codes not being detected by DECOTA. For example, there were select ‘silent’ codes not detected by DECOTA that were detected by humans in Dataset 3 – yet these all individually represented only > 0.1% of the data. Given this very low prevalence, DECOTA would not seek out these codes, making it a less suitable method when very uncommon themes are as important as common ones (e.g., in understanding the views of specific minority groups that may not be well represented in the data). Given these challenges with using STM, future work may seek to use a different Natural Language Processing approach to initially cluster topics.

Further limitations lie with our novel topic choice algorithm. Though our algorithm generally suggests an optimal value for K, it cannot use context about the use case to alter the decision, as humans may. As discussed by (Chang et al., 2009), there are instances in which the best number of topics in terms of model fit does not exactly match the best choice for K, for example when the user desires a smaller topic number for maximum interpretability. This occurred in Dataset 1, where our algorithm chose the topic number with the best model fit metrics. Using human judgement, it becomes clear that the second-best model fit indices (with fewer topics), would have aided interpretation of the topics and reduced repetition. Despite this, some detail would have been lost, especially for less prevalent codes. We therefore designed the model in a way that ensured users can manually alter the chosen topic number, if this circumstance was to arise.

Despite its high similarity to human coding, speed, and cost reduction – there will understandably be hesitancy in using tools like DECOTA for free-text analysis. This may be due to privacy concerns, biases in training data, and a potential lack of transparency. In our case, privacy concerns are somewhat mitigated by using the LLM’s API, which does not store participant data for use in training future models (Open AI, 2024a). In addition, all free-text data run through the API is anonymised.

Despite this, future work should seek to bring all data processing offline, so no free-text data would be shared with a third party. On biases, it is well documented that AI systems, such as chat bots, can develop undesirable human biases such as gender or racial stereotyping (Beattie et al., 2022). With our use case of summarising data, this concern is less pressing. However, OpenAI have made efforts to reduce biases by training their LLMs on diverse data sources and conducting adversarial testing, where models are intentionally prompted to generate outputs that might reveal biases, to then address. In our fine-tuning, we also mitigated this issue by carefully generating training data that was on a diverse range of topics, sentiments, and did not exhibit stereotypes. We also set the randomness (‘temperature’) of our LLM to zero to encourage more repeatable responses, based on phrase frequency rather than interpretation. Notably, this issue is not unique to DECOTA, as humans also have biases when conducting qualitative analyses. Finally, we attempted to be as transparent as possible in our analysis, by allowing users to see the results from each of the six methodological steps, edit values such as topic number choice and cut-off values when clustering the final themes, and providing a comprehensive output table with example quotes for each code.

5. Conclusion

For the first time, we propose and validate the Deep Computational Text Analyser (DECOTA) – the first systematically validated, open-access Large Language Model (LLM) tool to automatically analyse qualitative data. We demonstrate that it is an efficient, cost-effective, and rapid way to analyse large volumes of free-text data. Using a formal triangulation, we show that DECOTA yielded 91.6% of all ‘codes’ and 90% of ‘themes’ detected by a human inductive Thematic Analysis, but approximately 378 times faster and 1920 times cheaper. Compared to current computational methods such as Structural Topic Modelling (STM), we show that DECOTA is approximately 240 times faster and 1220 times cheaper – making it a significant step in automatically analysing qualitative data. Though further validation is required, our tool has potentially vast implications both within academia (e.g., in developing new psychometric measures) and beyond, such as in understanding the public’s initial perceptions about a proposed policy. More broadly, DECOTA shows that it is possible to use NLP and LLMs to transparently and rigorously analyse free-text data.

6. References

Alarcon, N. (2020, July 7). *OpenAI Presents GPT-3, a 175 Billion Parameters Language Model*. NVIDIA Technical Blog. https://developer.nvidia.com/blog/openai-presents-gpt-3-a-175-billion-parameters-language-model/

Amazon. (2023, December 13). *How Amazon Continues To Improve The Customer Reviews Experience With Generative AI*. https://www.aboutamazon.co.uk/news/innovation-at-amazon/customer-reviews-generative-ai

AWS. (2024). *What are Large Language Models? - LLM AI Explained - AWS*. Amazon Web Services, Inc. https://aws.amazon.com/what-is/large-language-model/

Baker, P. (2023). *Using Corpora in Discourse Analysis*. Bloomsbury Publishing.

Beattie, H., Watkins, L., Robinson, W. H., Rubin, A., & Watkins, S. (2022). Measuring and Mitigating Bias in AI-Chatbots. *2022 IEEE International Conference on Assured Autonomy (ICAA)*, 117–123. https://doi.org/10.1109/ICAA52185.2022.00023

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brown, J. (2018). Interviews, Focus Groups, and Delphi techniques. In *Advanced Research Methods for Applied Psychology*. Routledge.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners* (arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165

Bua, A., & Escobar, O. (2018). Participatory-Deliberative Processes And Public Policy Agendas: Lessons For Policy And Practice. *Policy Design and Practice*, *1*(2), 126–140. https://doi.org/10.1080/25741292.2018.1469242

Byrne, D. (2022). A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & Quantity*, *56*(3), 1391–1412. https://doi.org/10.1007/s11135-021-01182-y

Caplena. (2024). *Analyze free text comments through AI | Caplena*. https://caplena.com/en/

Carter, N., Bryant-Lukosius, D., DiCenso, A., Blythe, J., & Neville, A. J. (2014). The Use Of Triangulation In Qualitative Research. *Oncology Nursing Forum*, *41*(5), 545–547. https://doi.org/10.1188/14.ONF.545-547

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J., & Blei, D. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems*, *22*. https://proceedings.neurips.cc/paper_files/paper/2009/hash/f92586a25bb3145facd64ab20fd554ff-Abstract.html

Chung, G., Rodriguez, M., Lanier, P., & Gibbs, D. (2022). Text-Mining Open-Ended Survey Responses Using Structural Topic Modeling: A Practical Demonstration to Understand Parents' Coping Methods During the COVID-19 Pandemic in Singapore. *Journal of Technology in Human Services*, *40*(4), 296–318. https://doi.org/10.1080/15228835.2022.2036301

Church, K. W., Chen, Z., & Ma, Y. (2021). Emerging Trends: A Gentle Introduction To Fine-Tuning. *Natural Language Engineering*, *27*(6), 763–778. https://doi.org/10.1017/S1351324921000322

Crowston, K., Allen, E. E., & Heckman, R. (2012). Using Natural Language Processing Technology For Qualitative Data Analysis. *International Journal of Social Research Methodology*, *15*(6), 523–543. https://doi.org/10.1080/13645579.2011.625764

del Gobbo, E., Fontanella, S., Sarra, A., & Fontanella, L. (2021). Emerging Topics in Brexit Debate on Twitter Around the Deadlines. *Social Indicators Research*, *156*(2), 669–688. https://doi.org/10.1007/s11205-020-02442-4

Department for Transport. (2024). *Public Attitudes To The Use Of AI In Dft Consultations And Correspondence*. https://www.gov.uk/government/publications/public-attitudes-to-the-use-of-ai-in-dft-consultations-and-correspondence

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Efimov, V. (2023, September 16). *Large Language Models: SBERT — Sentence-BERT*. Medium. https://towardsdatascience.com/sbert-deb3d4aef8a4

Ferrer, J. (2024). *An Introductory Guide to Fine-Tuning LLMs*. https://www.datacamp.com/tutorial/fine-tuning-large-language-models

FHIRFLY. (2023, August 24). Tuning the Temperature of GPT Models: A Guide to Reducing Hallucinations. *Medium*. https://medium.com/@fhirfly/tuning-the-temperature-of-gpt-models-a-guide-to-reducing-hallucinations-a92af019c534

Fugard, A. J. B., & Potts, H. W. W. (2015). Supporting Thinking On Sample Sizes For Thematic Analyses: A Quantitative Tool. *International Journal of Social Research Methodology*, *18*(6), 669–684. https://doi.org/10.1080/13645579.2015.1005453

IBM. (2024a). *What Are Large Language Models (LLMs)?* IBM. https://www.ibm.com/topics/large-language-models

IBM. (2024b). *What are AI Hallcuinations?* IBM. https://www.ibm.com/topics/ai-hallucinations

Kim, S., Lee, N., & King, P. (2020). Dimensions of Religion and Spirituality: A Longitudinal Topic Modeling Approach. *Journal for the Scientific Study of Religion*, *59*. https://doi.org/10.1111/jssr.12639

Kong, H.-J. (2019). *Managing Unstructured Big Data in Healthcare System*. https://e-hir.org/journal/view.php?doi=10.4258/hir.2019.25.1.1

Lee, A. (2023, January 26). *What Are Large Language Models Used For?* NVIDIA Blog. https://blogs.nvidia.com/blog/what-are-large-language-models-used-for/

Lee, T. Y., Smith, A., Seppi, K., Elmqvist, N., Boyd-Graber, J., & Findlater, L. (2017). The Human Touch: How Non-Expert Users Perceive, Interpret, And Fix Topic Models. *International Journal of Human-Computer Studies*, *105*, 28–42. https://doi.org/10.1016/j.ijhcs.2017.03.007

Lindstedt, N. C. (2019). Structural Topic Modeling For Social Scientists: A Brief Case Study with Social Movement Studies Literature, 2005–2017. *Social Currents*, *6*(4), 307–318. https://doi.org/10.1177/2329496519846505

Manifesto Club. (2023, February 7). *LTNs And The Side Stepping Of Public Opinion*. https://manifestoclub.info/ltns-and-the-side-stepping-of-public-opinion/

McHugh, M. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica : Časopis Hrvatskoga Društva Medicinskih Biokemičara / HDMB*, *22*, 276–282. https://doi.org/10.11613/BM.2012.031

Meaney, C., Stukel, T. A., Austin, P. C., Moineddin, R., Greiver, M., & Escobar, M. (2023). Quality Indices For Topic Model Selection And Evaluation: A Literature Review And Case Study. *BMC Medical Informatics and Decision Making*, *23*(1), 132. https://doi.org/10.1186/s12911-023-02216-1

Mickelsson, J., van Haren, J. J. G. M., & Lemmink, J. G. A. M. (2022). Wrinkles in a CSR Story: Mismatched Agendas In Fast Fashion Service Brands’ CSR Reputation. *Journal of Service Management*, *34*(2), 256–273. https://doi.org/10.1108/JOSM-07-2021-0243

Mimno, D., & Lee, M. (2014). Low-dimensional Embeddings for Interpretable Anchor-based Topic Inference. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1319–1328). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1138

Moser, A., & Korstjens, I. (2017). Series: Practical Guidance To Qualitative Research. Part 1: Introduction. *European Journal of General Practice*, *23*(1), 271–273. https://doi.org/10.1080/13814788.2017.1375093

Mostafa, M. M. (2023). A One-Hundred-Year Structural Topic Modeling Analysis Of The Knowledge Structure Of International Management Research. *Quality & Quantity*, *57*(4), 3905–3935. https://doi.org/10.1007/s11135-022-01548-w

Mu, W., Lim, K. H., Liu, J., Karunasekera, S., Falzon, L., & Harwood, A. (2022). A Clustering-Based Topic Model Using Word Networks And Word Embeddings. *Journal of Big Data*, *9*(1), 38. https://doi.org/10.1186/s40537-022-00585-4

Naeem, M., Ozuem, W., Howell, K., & Ranfagni, S. (2023). A Step-by-Step Process of Thematic Analysis to Develop a Conceptual Model in Qualitative Research. *International Journal of Qualitative Methods*, *22*, 16094069231205789. https://doi.org/10.1177/16094069231205789

National Audit Office. (2019). *Challenges In Using Data Across Government*. https://www.nao.org.uk/insights/challenges-in-using-data-across-government/

Nelson, L. K., Burk, D., Knudsen, M., & McCall, L. (2021). The Future of Coding: A Comparison of Hand-Coding and Three Types of Computer-Assisted Text Analysis Methods. *Sociological Methods & Research*, *50*(1), 202–237. https://doi.org/10.1177/0049124118769114

Nikolenko, S. I., Koltcov, S., & Koltsova, O. (2017). *Topic Modelling For Qualitative Studies*. https://journals.sagepub.com/doi/abs/10.1177/0165551515617393

Open AI. (2023). *GPT-3.5 Turbo Fine-Tuning And API Updates.* OpenAI. https://openai.com/index/gpt-3-5-turbo-fine-tuning-and-api-updates/

Open AI. (2024a). *Enterprise Privacy*. https://openai.com/enterprise-privacy/

Open AI. (2024b). *OpenAI Platform*. https://platform.openai.com

Pandur, M. B., Dobša, J., & Kronegger, L. (2020). Topic Modelling in Social Sciences—Case Study of Web of Science*. *Central European Conference on Information and Intelligent Systems*, 211–218. https://www.proquest.com/docview/2531367517/abstract/56E58F23EA6B448CPQ/1

Pisaneschi, B. (2024). *Unstructured Data and AI: Fine-Tuning LLMs to Enhance the Investment Process*. https://doi.org/10.56227/24.1.9

Povitkina, M., Carlsson Jagers, S., Matti, S., & Martinsson, J. (2021). Why Are Carbon Taxes Unfair? Disentangling Public Perceptions Of Fairness. *Global Environmental Change*, *70*, 102356. https://doi.org/10.1016/j.gloenvcha.2021.102356

Rich, J. L., Chojenta, C., & Loxton, D. (2013). Quality, Rigour and Usefulness of Free-Text Comments Collected by a Large Population Based Longitudinal Study—ALSWH. *PLoS ONE*, *8*(7), e68832. https://doi.org/10.1371/journal.pone.0068832

Tenny, S., Brannan, J. M., & Brannan, G. D. (2024). Qualitative Study. In *StatPearls*. StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK470395/

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large Language Models In Medicine. *Nature Medicine*, *29*(8), 1930–1940. https://doi.org/10.1038/s41591-023-02448-8

Towler, L., Bondaronek, P., Papakonstantinou, T., Amlôt, R., Chadborn, T., Ainsworth, B., & Yardley, L. (2023). Applying Machine-Learning To Rapidly Analyze Large Qualitative Text Datasets To Inform The COVID-19 Pandemic Response: Comparing Human And Machine-

Assisted Topic Analysis Techniques. *Frontiers in Public Health*, *11*, 1268223. https://doi.org/10.3389/fpubh.2023.1268223

Ünver, A., & Kurnaz, A. (2022). Securitization of Disinformation in NATO's Lexicon: A Computational Text Analysis. *All Azimuth: A Journal of Foreign Policy and Peace*, *11*(2), Article 2. https://doi.org/10.20991/allazimuth.1110500

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wall, T. (2020, September 20). The New Road Rage: Bitter Rows Break Out Over UK's Low-Traffic Neighbourhoods. *The Observer*. https://www.theguardian.com/world/2020/sep/20/the-new-road-rage-bitter-rows-break-out-over-uks-low-traffic-neighbourhoods

Weston, S. J., Shryock, I., Light, R., & Fisher, P. A. (2023). Selecting the Number and Labels of Topics in Topic Modeling: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *6*(2), 25152459231160105. https://doi.org/10.1177/25152459231160105

Wright, L., Fluharty, M., Steptoe, A., & Fancourt, D. (2022). How Did People Cope During the COVID-19 Pandemic? A Structural Topic Modelling Analysis of Free-Text Data From 11,000 United Kingdom Adults. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.810655

Xiang, L. (2022). Evaluation of Manuscript Preference of Sci-Tech Periodicals: Structural Topic Models of Computer Engineering Related Publications in China. *SCIREA Journal of Electrical Engineering*, *7*(2), 22–43. https://doi.org/10.54647/dee47268

Zade, H., Drouhard, M., Chinh, B., Gan, L., & Aragon, C. (2018). Conceptualizing Disagreement in Qualitative Coding. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–11. https://doi.org/10.1145/3173574.3173733