
Proceedings Paper:
Mu, Y., Dong, C., Bontcheva, K. orcid.org/0000-0001-6152-9600 et al. (1 more author) (2024) Large language models offer an alternative to the traditional approach of topic modelling. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 20-25 May 2024, Torino, Italy. ELRA and ICCL , pp. 10160-10171. ISBN 978-2-493814-10-4

Large Language Models Offer an Alternative to the Traditional Approach of Topic Modelling

Yida Mu, Chun Dong, Kalina Bontcheva, Xingyi Song
Department of Computer Science, The University of Sheffield
{y.mu, cdong8, k.bontcheva, x.song}@sheffield.ac.uk

Abstract

Topic modelling, as a well-established unsupervised technique, has found extensive use in automatically detecting significant topics within a corpus of documents. However, classic topic modelling approaches (e.g., LDA) have certain drawbacks, such as the lack of semantic understanding and the presence of overlapping topics. In this work, we investigate the untapped potential of large language models (LLMs) as an alternative for uncovering the underlying topics within extensive text corpora. To this end, we introduce a framework that prompts LLMs to generate topics from a given set of documents and establish evaluation protocols to assess the clustering efficacy of LLMs. Our findings indicate that LLMs with appropriate prompts can stand out as a viable alternative, capable of generating relevant topic titles and adhering to human guidelines to refine and merge topics. Through in-depth experiments and evaluation, we summarise the advantages and constraints of employing LLMs in topic extraction.

Keywords: Large Language Models, Topic Modelling, LLM-driven Topic Extraction, Evaluation Protocol

1. Introduction

Understanding the topics within a collection of documents is crucial for various academic, business, and research disciplines (Ramage et al., 2009; Vayansky and Kumar, 2020). Gaining insights into the primary topics can help in organising, summarising, and drawing meaningful conclusions from vast amounts of textual data.

Classic approaches to topic analysis including 1) **topic modelling**: an unsupervised approach used to identify themes or topics within a large corpus of text by analysing the patterns of word occurrences (Blei et al., 2003; Grootendorst, 2022); and 2) **close-set topic classification**: model trained on sufficient labelled data with pre-defined close-set topics (Wang and Manning, 2012; Song et al., 2021; Antypas et al., 2022). However, these approaches have certain limitations and challenges. Topic classification requires a predefined, closed set of topics and is unable to capture unseen topics. While topic modelling might produce very broad topics while missing out on nuanced or more specific sub-topics that might be of interest (i.e., topic granularity) (Abdelrazek et al., 2023). Besides, the topics generated by models such as LDA and BERTopic are clusters of words with associated probabilities. Sometimes these clusters might not make intuitive sense to human interpreters, leading to potential misinterpretations (Gillings and Hardie, 2023).

In addition, these approaches do not perform well on handling unseen documents without a complete re-run of the model, which consequently makes it less efficient for dynamic datasets that are frequently updated (e.g., a dynamic Twitter corpus) (Blei and Lafferty, 2006; Wang et al., 2008).

Addressed those limitations, we proposed an alternative topic modelling approach in this paper – **Topic Extraction using Large Generative Language Models**

Generative transformer-based large language models (LLMs) (Vaswani et al., 2017), such as GPT (Brown et al., 2020) and LLaMA (Touvron et al., 2023a,b), have obtained significant attention for their proficiency in understanding and generating human-like languages. Prompt-based LLMs are transforming conventional natural language processing (NLP) workflows (Brown et al., 2020) from model training to evaluation protocols. For example, existing LLMs (e.g., GPT-4) trained using reinforcement learning with human feedback (RLHF) (Ziegler et al., 2019; Ouyang et al., 2022) has shown compatible zero-shot classification performance against supervised methods (e.g., a fully fine-tuned BERT (Devlin et al., 2019)) in various natural language understanding tasks (e.g., detecting customer complaints) in computational social science (Ziems et al., 2023; Mu et al., 2023c).

Due to their plug-and-play convenience, LLMs bring transformative potential to topic modelling. Given that LLMs have the strong capabilities of zero-shot text summarising on par with human annotators (Zhang et al., 2024), we argue LLMs might leverage their inherent understanding of language nuances to extract (or generate) topics. Their ability to comprehend context, nuances, and even subtle thematic undertones has been demonstrated in various NLP tasks, allowing for a richer and more detailed categorisation of topics (Wu et al., 2023; Tang et al., 2023). Besides, the prompt-based model inference pipeline allows users to add manual instructions to guide the model in generating customised outputs (Ouyang et al., 2022). Furthermore, LLMs

LREC-COLING 2024, pages 10160–10171
20-25 May, 2024.

can seamlessly adapt to evolving language trends and emerging topics (e.g., streaming posts on Twitter), ensuring that topic modelling remains relevant and up-to-date.

Given the lack of prior work, we shed light on the following research aims ranging from model inference and evaluation protocol:

- (i) Investigate the suitability of LLMs as a straightforward, plug-and-play tool for topic extraction without the necessity of complex prompts.
- (ii) Identify and address the limitations and challenges encountered in utilising LLMs for topic extraction.
- (iii) Assess the ability of LLMs to consistently adhere to human-specified guidelines in generating topics with desired granularity.
- (iv) Develop an evaluation protocol to measure the quality of topics generated by LLMs.

To this end, we make the following contributions:

- By conducting a series of progressive experiments[1] using different sets of prompting and manual rules, we observed that LLMs with appropriate prompts can be a strong alternative to traditional approaches of topic modelling.
- We empirically show that LLMs are capable of not just generating topics but also condensing overarching topics from their outputs. The resulting topics, complete with explanations, are easily understood by humans.
- We introduce evaluation metrics to assess the quality of topics organically produced by LLMs. These metrics are suitable for labelled or unlabelled datasets.
- Finally, a case study is provided to show the applications of LLMs in real-world scenarios (e.g., analysing topic trends over time). We demonstrate that LLMs can independently perform topic extraction and generate explanations for analysing temporal corpus from a dynamic Twitter dataset (See Figure 4).

2. Related Work

2.1. Topic Modelling

Topic modelling, as a classic unsupervised machine learning approach in computer science, has been broadly employed in various fields such as social science and bio-informatics for processing large-scale of documents (Blei et al., 2003; Song et al., 2021; Grootendorst, 2022). A standard output of a topic modelling algorithm is a set of fixed or flexible numbers of topics, where each topic can be typically represented by a list of top words. One can use manual or automatic methods to interpret a topic with corresponding top tokens (e.g., assign a meaningful name for each topic) (Lau et al., 2010; Allahyari and Kochut, 2015). However, topic interpretation is not always straightforward (Aletras and Stevenson, 2014). For example, the practice of assigning labels through an eyeballing approach often leads to incomplete or incorrect topic labels (Gillings and Hardie, 2023). Furthermore, topic labelling and interpretation rely heavily on the specialised knowledge of annotators (Lee et al., 2017). Besides, preprocessing (e.g., stemming and lemmatisation) can significantly affect topic modelling performance (Chuang et al., 2015; Schofield and Mimno, 2016). Therefore, the use of topic modelling often requires text pre-processing on the input documents and post-processing on the model outputs (e.g., topic labelling) to make the results human-interpretable (Vayansky and Kumar, 2020).

2.2. Close-set Topic Classification

On the other hand, closed-set topic classification serves as an alternative to unsupervised topic modelling approaches, which usually depend on models trained on datasets with predefined labels. Topic classification approaches have been widely applied on various domains such as computational social science (Wang and Manning, 2012; Iman et al., 2017) and biomedical literature categorisation (Lee et al., 2006; Stepanov et al., 2023). For example, during the COVID pandemic, topic classification approaches were used to analyse the spread of COVID-related misinformation (Song et al., 2021) and public attitudes towards vaccination (Poddar et al., 2022; Mu et al., 2023a). However, given the nature of the supervised classification task, topic classification typically requires a high cost of human effort in data annotation (Antypas et al., 2022). Meanwhile, in the context of labelling social media posts, predefined topics may overlap (such as 'News' and 'Sports'), leading to disagreements among annotators (Antypas et al., 2022).

2.3. LLMs-driven Topic Extraction

LLMs have demonstrated their capabilities in text summarisation tasks across various domains, such as news, biomedical, and scientific articles (Wu et al., 2023; Shen et al., 2023; Tang et al., 2023). Extractive text summarisation methods can precede LLM-driven topic extraction, i.e., simplifying document complexity and concentrating the topic

[1] Our source code: https://github.com/GateNLP/LLMs-for-Topic-Modeling

extraction on the most relevant content. (Srivastava et al., 2022; Joshi et al., 2023).

Meanwhile, LLMs also complement topic modelling approaches, reducing the need for human involvement in the interpretation and evaluation of topics. Stammbach et al. (2023) explore the use of LLMs for topic evaluation, uncovering that vanilla LLMs (e.g., ChatGPT) can be used as an out-of-the-box approach to automatically assess the coherence of topic word collections. By comparing the human and machine-produced interpretations, (Rijcken et al., 2023) uncover that LLMs ratings highly correlate with human annotations. Besides, topics generated by LLMs are more preferred by general users than the original categories (Li et al., 2023). Wang et al. (2023) and Xie et al. (2021) point out that LLMs are implicitly topic models which can be used to identify task-related information from demonstrations.

2.4. Our Work

In general, previous work has predominantly centred on utilising LLMs as assistants to enhance topic modelling approaches (e.g., automatic evaluation and topic labelling). These studies primarily rely on the output from topic modelling approaches like LDA and BERTopic, instead of topics directly generated by LLMs. In this work, we shed light on the potential of using **LLMs exclusively** for topic extraction and assess topics generated by LLMs from scratch, which is a different task compared to topic modelling and closed-set topic classification.

3. Models and Datasets

3.1. LLMs

In this study, we assess the capability of two widely used LLMs in topic extraction.

- **GPT-3.5 (GPT)**[2] represents an advanced iteration of the GPT-3 language model, enhanced with instruction fine-tuning. Through the OpenAI API, GPT offers plug-and-play capabilities for numerous NLP tasks such as machine translation, common sense reasoning, and question & answering.
- **LLaMA-2-7B (LLAMA)** (Touvron et al., 2023b) is an enhanced iteration of LLaMA 1 (Touvron et al., 2023a), trained on a corpus that is 40% larger and with twice the context length. We employ the LLaMA model through the Hugging Face platform[3] (Wolf et al., 2020).

We chose GPT and LLaMA as they represent two primary modes of LLMs: API-based commercial product and fine-tunable open-source model. Both LLMs have been frequently selected as base models in prior LLM evaluation studies (Ziems et al., 2023; Mu et al., 2023c). Note that there are stronger alternatives such as the GPT-4 and LLaMA-2-70B. However, the chosen models offer more practical implications in terms of financial considerations and computational resources, such as the number of GPUs required and API pricing.

For comparison, we also compare to two widely used baseline models, namely LDA (Blei et al., 2003) and BERTopic (Grootendorst, 2022). We utilise LLMs to generate final topic names based on a list of tokens for each topic.

3.2. Datasets

To assess the generalisability of LLMs, we examine one open-domain dataset and one domain-specific dataset. We chose these two datasets because they contain texts of varying lengths (i.e., document v.s. sentence levels) and density of vocabulary (i.e., diverse v.s. similar vocabularies). Note that topic modelling approaches might struggle with very short texts (e.g., user-generated content on social media) due to the lack of sufficient context to derive meaningful topics. Besides, the Twitter dataset also provides temporal information which can be used for analysis topics trends over time (See Case Study in **§** 5).

- **20 News Group (20NG)** (Lang, 1995), as a classic benchmark[4], has been widely used in various NLP downstream tasks such as text classification and clustering.
- **CAVS (VAXX)**[5] (Poddar et al., 2022) is a fine-grained Twitter dataset designed for analysing reasons behind COVID-19 vaccine hesitancy. It contains a collection of tweets labelled under one of ten major vaccine hesitancy categories, such as *'Side-effect' and 'Vaccine Ineffective'*.

Pre-processing For the Twitter dataset, we perform standard text cleaning rules to filter out all user mentions (i.e., @USER) and hyperlinks. We employ a stratified data split method to sample 20% of documents from each dataset for the test set, maintaining the same category ratios as in the original dataset.

4. Experiments

In this section, we outline our experimental setup, covering both prompt engineering and the strate-

[2] https://platform.openai.com/docs/models/gpt-3-5

[3] https://huggingface.co/meta-llama/Llama-2-7b-chat

[4] http://qwone.com/~jason/20Newsgroups/

[5] https://github.com/sohampoddar26/caves-data

gies we adopted to improve the generation of expected topics. Given the nascent nature of the task, we structure our experiments in a sequence from simpler to more complex prompting settings. This incremental experimental approach aids us in identifying challenges as they arise, guiding us to devise appropriate solutions.

4.1. Experiment 1: Out-of-box (Basic Prompt)

We first explore the use of LLMs as an out-of-box approach for topic extraction. Due to the quadratic complexity of the transformer architecture's attention mechanism with respect to the input sequence length (Vaswani et al., 2017), LLMs struggle to summarise topics from a large corpus in one prompt. For instance, even the latest GPT-4[6], which has expanded its maximum input limit to 32,000 tokens (equivalent to approximately 25,000 words in English), is still incapable of processing most NLP datasets in a single pass.

4.1.1. Prompting Strategies

Consequently, we investigate two prompting strategies: **(i) feeding text individually and (ii) feeding in batched text** (e.g., 20 documents per batch). Note that the former approach incurs slightly higher costs as it necessitates a full prompting message with each iteration.

As illustrated in Figure 1, our prompt comprises two parts: (i) a system prompt to help the model understand human instructions and the desired output format, and (ii) a user prompt to provide the documents for topic extraction. A structured output format is crucial for subsequent topic statistics and evaluation.

4.1.2. Results and Discussion

Given the large number of topics obtained, we count the number of occurrences of each topic and then list the top K topics from the final list. The count of the number of topics represents the proportion of each topic in a given dataset. However, from our initial set of experiments, we observe that both prompting strategies result in similar results. We empirically find that LLMs can reliably handle up to 20 documents per pass depending on the average length.

According to the results of both GPT and LLaMA with 'Out-of-box' (See Table 1, GPT & LLaMA Expt. 1 Basic Prompt), we observe that LLMs struggle to produce quality topics when given basic instructions without any manual constraints. Using the VAXX dataset (i.e., fine-grained reasons related to vaccine hesitancy) as an example, we identify the following challenges:

- **Challenge (i)** GPT tends to produce very general topics like *'Vaccine', 'COVID Vaccination' and 'Vaccine Hesitancy'*, which are already apparent as the primary themes of the dataset. It suggests that GPT without further manual instructions may not understand the granularity of topics we expected. Note that LLaMA did not return such general with this setting.

- **Challenge (ii)** By manually examining the list of final topics, we also observe a significant overlap in the topics generated by both LLMs, with many topics essentially conveying the same meaning. For example, LLMs might generate topics in various cases and formats, such as *'side-effect', 'Side Effect', 'serious side effect', 'fear of side effects' and 'vaccine side effect'.*

- **Challenge (iii)** Consequently, LLMs generate a large list of topics, which poses a significant challenge in selecting representative topics. From the VAXX dataset, both LLMs return

Basic Prompt = '''
<s>[INST] <<SYS>>
Read the text below and list up to 3 topics. Each topic should contain fewer than 3 words. Ensure you only return the topic and nothing more. \n
The desired output format:
Topic 1: xxx\nTopic 2: xxx\nTopic 3: xxx
<</SYS>>
{text} **[/INST]'''**

Basic Prompt + Seeds Topic= '''
<s>[INST] <<SYS>>
Consider the previous extracted topics:
{existing_topics}
Read the text below and list up to 3 topics. Each topic should contain fewer than 3 words. Ensure you only return the topic and nothing more. \n The desired output format:
Topic 1: xxx\nTopic 2: xxx\nTopic 3: xxx
<</SYS>>
{text} **[/INST]'''**

Prompt for Summarization = '''
<s>[INST] <<SYS>>
Summarize and merge the following list of topics into {Fixed_Number} final topics.
<</SYS>>
{List_of_Topics} **[/INST]'''**

Figure 1: Example prompts for LLaMA. Text enclosed by the special tokens '≪*SYS*≫' is designated as a system prompt.

[6] https://openai.com/research/gpt-4

around 2,500 extracted topics, of which 60% are unique. Given the two existing challenges, simply selecting the top K most frequent topics from the output list of 'Experiment Expt. 1 Basic' (See Table 1) may not yield a truly representative set of topics.

4.1.3. Solutions

The initial challenges primarily concern obtaining high-quality topics, which are crucial for subsequent topic selection. To this end, we propose the following solutions:

- **Adding Constraints to the Prompt:** To avoid the overly broad topics generated by LLMs, we introduce additional constraints in the prompt. For instance, we guide the model not to return broader topics such as 'COVID-19' and 'COVID-19 Vaccine' in the system prompt. Besides, we also provide task specific information to guide the model to understand the granularity of the given dataset, e.g., by prompting LLMs to return topics related to COVID-19 vaccine hesitancy reasons.
- **Hand-crafted Rules:** We transform similar outputs by adding post-process rules after each iteration. These include using regular expressions to convert all words to lowercase and replace any 'Hyphens' with an 'Empty Space'. We also apply text lemmatisation rules to normalise all raw outputs (e.g., 'vaccine effectiveness' and 'vaccine effective').
- **Top K Topics** To identify representative topics within the given datasets, we start by employing a simple Top K method, focusing on topics that exhibit the highest frequencies. Further methods for selecting representative topics are discussed in Section § 4.3.

By integrating our proposed solutions into prompts, we achieve improved topic results (see Table 1, 'GPT & LLaMA Expt. 1 + Manual Instructions') compared to those from 'GPT & LLaMA Expt. 1 + Basic Prompt'.

4.2. Experiment 2: Topics Granularity (GPT & LLaMA Expt. 2 + Seeds Topic)

Topic modelling approaches can control the granularity of topics by setting the specific hyperparameter[7] to fix the number of topics. One can also conduct topic modelling with minimal domain knowledge by adding several anchor words (Gallagher et al., 2017).

Figure 2: Two examples of final topics summarised by LLMs, namely 'Technology & Computers' (20NG) and 'Trust & Mistrust' (Vaccine dataset).

As shown in Table 1, the solutions we proposed in experiment 1 can effectively filter out irrelevant topics. However, they fall short in addressing the more complex scenario of similar topics, such as 'vaccine side effect', 'fear side effect' and 'serious side effect'. Leveraging advanced capabilities in natural language understanding, we test an enhanced prompting setup by using seed topics. The purpose of providing seed topics is to guide the model towards discerning the granularity of the topics we anticipate. This is similar to how a human can understand the potential topics of an unseen set of documents by manually reviewing a few examples to get prior knowledge.

[7]For example, train a LDA via scikit-learn: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html

#	20 News Group
Original Categories	*Comp. [graphics, os.ms-windows.misc, sys.ibm.pc.hardware, sys.mac.hardware, windows.x, misc.forsale], Rec. [autos, motorcycles, sport.baseball, sport.hockey], Sci. [electronics, medical, space, crypt], Soc. [religion.christian], Talk. [politics.guns, politics.mideast, politics.misc, religion.misc], alt.atheism*
LDA	Lasting Impressions, People's Perception, Used Cars, New Knowledge, Problem-solving techniques, Armenian Genocide, New System, File Management
BERTopic	Game, God, Magnetism, Fire, Depression, Car, Encryption, Server, Technology, Operating System
GPT Expt. 1 Basic Prompt	fbi, faith, baseball, lie, hockey, god, software, cloud, baptism, microsoft
GPT Expt. 1 + Manual Instructions	fbi, faith, baseball, lie, hockey, god, software, microsoft, image conversion, pc
GPT Expt. 2 + Seeds Topic	computer hardware, baseball, religion, hockey, genocide, software, encryption, christianity, faith, price
GPT Expt. 3 Summarisation	Technology & Computers, Sports, Religion & Philosophy, Government & Law, Media & Entertainment, Health & Medicine, Vehicles & Transportation, Society & Social Issues, History & Politics, Miscellaneous
LLaMA Expt. 1 Basic Prompt	technology, future, innovation, price, email, genocide, evidence, sin, books, bible
LLaMA Expt. 1 + Manual Instructions	technology, future, innovation, genocide, bible, god, software, graphics, game, encryption
LLaMA Expt. 2 + Seeds Topic	baseball, hardware, car, software, windows, player, god, government, christianity, hockey
LLaMA Expt. 3 +Summarisation	Technology & Computers, Religion & Spirituality, Society, Culture, & Human Rights, Communication & Media, Sports, Recreation, & Hobbies, Science & Research, Economics & Business, Government, Law, & Politics, Health & Wellbeing, Miscellaneous
#	**Vaccine Hesitancy Reasons**
Original Categories	Conspiracy, Country, Ingredients, Mandatory, Pharma, Political, Religious, Rushed, Side effect, Unnecessary
LDA	Vaccine Safety, COVID-19 Vaccination, Vaccine Safety, COVID vaccine, Vaccine effectiveness, COVID-19 Vaccine, Vaccine Allergies, Vaccine Efficacy, COVID-19 Vaccine, COVID-19 Vaccination
BERTopic	Vaccines, COVID-19, Government spending, COVID, Effectiveness, Untrustworthy, Love, Pandemic, Aging, Influenza, Pandemic, COVID
GPT Expt. 1 Basic Prompt	vaccine hesitancy, vaccine, vaccine effectiveness, vaccine safety, hesitancy, side effects, safety, trust, lack of trust, vaccine efficacy
GPT Expt. 1 + Manual Instructions	vaccine effectiveness, vaccine safety, side effect, trust, vaccine efficacy, misinformation, personal choice, conspiracy theory, fear
GPT Expt. 2 + Seeds Topic	side effect, ineffective, rushed, lack of trust, vaccine safety, vaccine efficacy, death, testing, natural immunity, dangerous, government
GPT Expt. 3 Summarisation	Vaccine Efficacy & Safety, Trust & Hesitancy, Misinformation & Beliefs, Government & Political Influence, Economic & Financial Concerns, Economic & Financial Concerns, Vaccine Development & Availability, Public Perception & Response, Legal & Ethical, Comparison & Alternative, Global & Societal Impacts
LLaMA Expt. 1 Basic Prompt	safety concerns, lack of trust, misinformation, personal beliefs, side effects, fear of side effects, trust issues, efficacy doubts, personal freedom, effectiveness doubts
LLaMA Expt. 1 + Manual Instructions	safety concerns, lack of trust, misinformation, personal beliefs, side effects, trust issues, efficacy doubts, personal freedom, long term effects, lack of information
LLaMA Expt. 2 + Seeds Topic	side effects, ineffective, safety concerns, lack of trust, efficacy doubts, personal beliefs, effectiveness, misinformation, long term effects, death
LLaMA Expt. 3 +Summarisation	Trust & Misinformation, Safety & Side Effects, Efficacy Doubts, Autonomy & Personal Beliefs, Economic & Corporate Concerns, Mandatory Vaccination Concerns, Political & Social Influences, Medical & Health Concerns, Access & Availability, Others

Table 1: For both LLMs and baseline models, we present the top 10 topics. Additionally, we include the original categories of each dataset for reference.

4.2.1. Prompting Strategies

We select two categories from the original list of labels to serve as seeds topic. They are injected into the prompts to guide the LLMs in understanding the granularity of the topics we anticipate. The example of the prompt demonstrated in the Figure 1, second row "Basic Prompt + Seeds Topic".

4.2.2. Results and Discussion

In Table 1, we note that incorporating seed topics (rows titled with Set2 + Seeds Topic) consistently enhances the performance of LLMs in topic extraction across various datasets and LLMs. This indicates that adding seed topics can guide the model in understanding the desired granularity of topics.

4.3. Experiment 3 Generating Final List (GPT & LLaMA Expt. 3 + Summarisation)

To obtain the final list of topics that can best represent a given set of documents, we consider a

further strategy to merge topics into N number of final topics:

Topic Summarisation We introduce an additional round of experiments which prompt LLMs to extract the N most appropriate topics from the extracted topic list. While the extracted list is expansive, it is still within the processing capacity of most LLMs, such as the 16k context length of GPT-3.5. For this experiment, we use all raw topics (i.e., the results of GPT & LLaMA Set 2 + Seed Topics) as input. Through specific prompts, we guide LLMs to produce easily interpretable final N topics with varying granularity. The end result of this process closely mirrors the output format of LDA and BERTopic, where each topic is accompanied by a list of subtopics.

Prompting Strategy To guide LLMs in summarising from the final list of topics, we use a prompt that directly asks the model to merge and summarise the given topic list. We also employ a few-shot prompting strategy by manually adding an example to guide the model in generating topics with the desired granularity. The example of the prompt demonstrated in the Figure 1, third row "Prompt for Summarisation".

Results and Discussion As indicated in Table 1 (GPT & LLaMA Expt. 3 + Summarisation), the final set of 10 topics encompasses the majority of the original categories from the source datasets. This highlights the potent proficiency of LLMs in summarising extensive corpora, as demonstrated in prior text summarisation tasks (Tang et al., 2023; Pu et al., 2023; Zhang et al., 2024). Moreover, LLMs offer explanations that are easily understandable by humans, detailing the content each topic encompasses. In Figure 2, we showcase examples illustrating how LLMs produce interpretable final topics derived from both datasets.

4.4. Topic Extraction Evaluation

Previous work has employed evaluation metrics such as perplexity and coherence score (Aletras and Stevenson, 2013). However, due to the new format of topics generated by LLMs, existing evaluation pipelines are unable to fully handle it. In Table 1, it is evident that the final Top N list produced by LLMs offers better granularity and interpretability than topics generated using the basic prompt. Nonetheless, having an automated evaluation protocol is crucial for an empirical comparison of model performance. We elucidate our proposed evaluation metrics using outputs from the vaccine dataset:

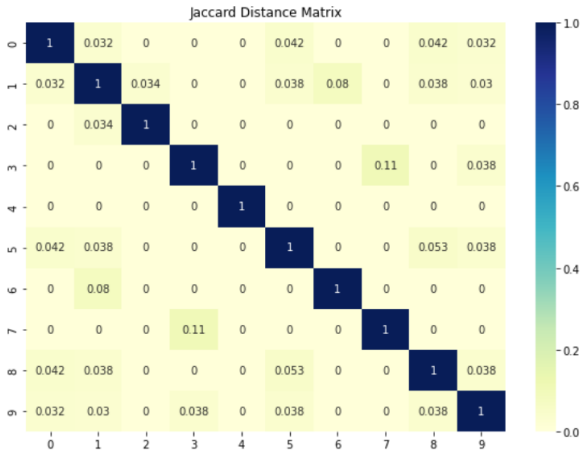

Figure 3: Matrix showing the Jaccard Distance between sub-topics derived from the top 10 general topic in the final list.

- **(i) Topic Distance over Top N Topics** We first use the Jaccard Distance to assess the sub-topics associated with the top N general topics. Given a set of N general topics, where each topic contains 10 sub-topics, we aim to compute the Jaccard distance between each pair of general topics. The Jaccard distance measures dissimilarity between two sets. For two sub-topic lists A and B, the Jaccard distance is defined as:

$$Jaccard\_distance(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

  where:

  - $|A \cap B|$ is the number of elements common to both lists A and B.
  - $|A \cup B|$ is the total number of unique elements across both lists A and B.
  - $Jaccard\_distance(A, B)$ ranges from 0 to 1, where 1 indicates that the lists are identical (i.e., all topics in A are in B and vice versa), and 0 indicates that the lists share no topics in common.

  Figure 3 shows that the topics in the final list are mostly distinct from one another (topics obtained from the LLaMA Expt. + Summarisation on the VAXX dataset).

- **(ii) Granularity of Top N Topics** We hypothesise that an increased number of topics results in decreased granularity (i.e., higher semantic similarity). To compute the average semantic similarity between each pair of topics from a final top N topics using the cosine similarity of BERT embeddings (Devlin et al., 2019), we define the following:

- Emb(T_i): BERT embedding (i.e., the '[CLS]' token) of the i-th topic (768 dimensions).
- Similar(T_i, T_j): Cosine similarity between the BERT embeddings of the i-th and j-th topics.
- N: Top N topics.

Given N topics, the average semantic similarity between each pair of topics can be computed as follows:

- Compute the BERT embeddings for each topic Emb(T_i) for $i = 1, 2, ..., N$.
- Calculate the cosine similarity Similar(T_i, T_j) for each unique pair (T_i, T_j), where $i \neq j$ and $i, j = 1, 2, ..., N$.
- Compute the average of these similarities obtained above:

$$\text{Ave.} = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \text{Similar}(T_i, T_j) \quad (2)$$

This equation ensures that each pair is considered only once, as similarity(T_i, T_j) = similarity(T_j, T_i), and there are $\frac{N(N-1)}{2}$ unique pairs among N topics. The factor of 2 in the numerator adjusts for the fact that we are considering each pair only once in the double summation.

For our task, we compute the average semantic similarity from the final top N topics, where N takes values of 10, 20, and 30. We notice a positive trend where the average semantic similarity rises with an increase in the number of top N final topics, i.e., Top 10 (0.155), Top 20 (0.197), and Top 30 (0.203). This suggests that LLMs are capable of effectively summarise fine-grained Top N topics when provided with an extensive list of topics.

- **(iii) Recall** Using the seed topics (ST) as a reference, we employ the 'Recall' metric to assess how effectively the model can generate pertinent topics (i.e., adhering to human instructions). The recall score is determined by calculating the ratio of correctly identified seed topics to the total number of examples originally labelled as one of the seed topics.

$$\text{Recall} = \frac{\text{No. Correct Extracted ST Samples}}{\text{No. Seeds Topic Samples}} \quad (3)$$

- **(iv) Precision** Similarly, we compute the precision by determining the ratio of correctly identified seed topics to the total number of examples labelled as a seed topic by LLMs.

$$\text{Precision} = \frac{\text{No. Correct Extracted ST Samples}}{\text{No. Samples ST Extracted}} \quad (4)$$

For the Vaccine dataset, we finally obtain a higher 'Recall' (70.0) and a lower 'Precision' (49.6) based on the results from the LLaMA Expt. 2 + Seeds Topics.

5. Case Study: Temporal Analysis of COVID-19 Vaccine Hesitancy

Understanding the shifting reasons for hesitancy towards the COVID-19 vaccine is vital, as this knowledge can assist policymakers and biomedical companies in gauging public reactions (Poddar et al., 2022; Mu et al., 2023b). As time changes, new events, such as emerging reasons for vaccine reluctance, appear that might not have been evident in previous datasets. This indicates that an established LDA or BERTopic model might struggle to process tweets containing these novel topics.

5.1. Experimental Setup

We explore the performance of LLMs in addressing unseen topics by processing text in the chronological order. For this purpose, we utilise the Vaccine dataset (Poddar et al., 2022), as timestamp information is provided in the Twitter metadata.

Following the timeline of COVID-19 vaccine development[8], we first arrange the Vaccine dataset in chronological order from oldest to latest. We then divide it into three periods: **(a)** Pre-COVID-19 (before January 2020), **(b)** the COVID-19 vaccine development period (January 2020 to December 2020), and **(c)** the period post the first jab of the COVID-19 vaccine when the vaccine became widely adopted globally (after December 2020).[9]

5.2. Results and Discussions

Figure 4 (top half) showcases the principal topics related to COVID-19 vaccine hesitancy reasons across these three time periods. We also present a description of the figure produced by the cutting-edge multi-model GPT-4 (bottom half). Our proposed pipeline illustrates that LLMs are capable of automatically executing topic extraction, visualisation (note that LLMs can also generate Python & R codes for visualising a given set of documents with statistics), and explanation (i.e., based on data visualisation figures). This indicates that both API-based and open source LLMs can serve as a robust substitute for what LDA and BERTopic offer to researchers from various disciplines.

[8] https://coronavirus.jhu.edu/vaccines

[9] https://www.rcn.org.uk/magazines/Bulletin/2020/Dec/May-Parsons-nurse-first-vaccine-COVID-19

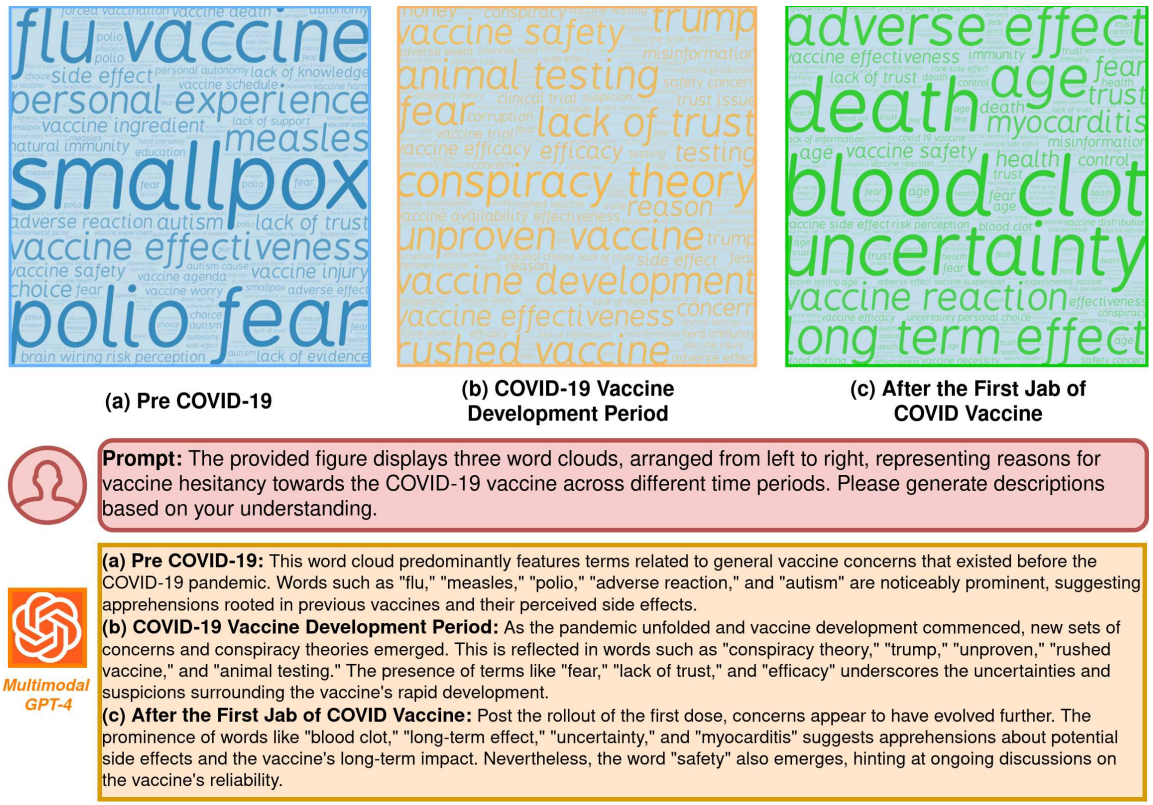

Figure 4: The figure showcases a trio of word clouds (upper half), each capturing the predominant reasons for vaccine hesitancy related to the COVID-19 vaccine over distinct time phases. Additionally, using the cutting-edge multi-modal GPT-4, we obtain descriptions (bottom half) derived from these word clouds.

6. Discussion

From the standpoint of practical implementation, we summarise the following main takeaways to address our proposed research questions:

- Owing to differences in the pre-training corpus and RLHF strategies, various LLMs can exhibit variability in 'zero-shot' topic extraction, especially when utilising only basic prompts.
- There is no 'one-size-fits-all' method of employing LLMs for topic extraction. We recommend conducting preliminary experiments on a small-scale test set. This approach helps early identify potential challenges or issues.
- Upon identifying these limitations, it becomes flexible to establish appropriate constraints and manual guidelines to assist LLMs in topic extraction. Given that LLMs have demonstrated their power in related tasks such as text summarisation and topic labelling, we argue that additional RLHF fine-tuning using a costumed dataset will bolster the LLMs' effectiveness in topic extraction.
- By incorporating seed topics, LLMs can generate topics with the desired granularity as specified by users.
- We propose several metrics to assess the quality of topics generated by LLMs from different perspectives, e.g., topic granularity.

7. Conclusion

In this work, we pioneer the exploration of utilising LLMs for topic extraction. Through empirical testing, we demonstrate that LLMs can serve as a viable and adaptable method for both topic extraction and topic summarisation, offering a fresh perspective in contrast to topic modelling methods. Additionally, LLMs demonstrate their capability to be directly applied to both specific-domain and open-domain datasets for topic extraction. This not only underlines the potential of LLMs in understanding hidden topics in large-scale corpora but also opens doors to various innovations (e.g., analysing dynamic datasets) in topic extraction.

In the future, we plan to concentrate on handling documents that surpass the maximum input length of current LLMs (e.g., LLaMA), for example, by extending the context window of LLMs (Chen et al., 2023; Peng et al., 2023). Additionally, we aim to develop new evaluation protocols to directly compare the results from topic modelling approaches and LLM-driven topic extraction, considering the distinct nature of the two tasks.

Ethics Statement

Our work has been approved by the Research Ethics Committee of our institute and complies with the policies of the Twitter API, the OpenAI API, and Meta LLaMA Terms&Conditions. All datasets are publicly available through the links provided in the original papers. All experiments using the OpenAI API cost less than 5 USD, which can be fully covered by the free trial credits.

Acknowledgements

This work has been funded by the UK's innovation agency (Innovate UK) grant 10039055 (approved under the Horizon Europe Programme as vera.ai[10], EU grant agreement 101070093).

References

Aly Abdelrazek, Yomna Eid, Eman Gawish, Walaa Medhat, and Ahmed Hassan. 2023. Topic modeling algorithms and applications: A survey. *Information Systems*, 112:102131.

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th international conference on computational semantics (IWCS 2013)–Long Papers*, pages 13–22.

Nikolaos Aletras and Mark Stevenson. 2014. Labelling topics using unsupervised graph-based methods. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 631–636.

Mehdi Allahyari and Krys Kochut. 2015. Automatic topic labeling using ontology-based topic models. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 259–264. IEEE.

Dimosthenis Antypas, Asahi Ushio, Jose Camacho-Collados, Vitor Silva, Leonardo Neves, and Francesco Barbieri. 2022. Twitter topic classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3386–3400.

David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.

Jason Chuang, Margaret E Roberts, Brandon M Stewart, Rebecca Weiss, Dustin Tingley, Justin Grimmer, and Jeffrey Heer. 2015. Topiccheck: Interactive alignment for assessing topic model stability. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–184.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186.

Ryan J Gallagher, Kyle Reing, David Kale, and Greg Ver Steeg. 2017. Anchored correlation explanation: Topic modeling with minimal domain knowledge. *Transactions of the Association for Computational Linguistics*, 5:529–542.

Mathew Gillings and Andrew Hardie. 2023. The interpretation of topic models for scholarly analysis: An evaluation and critique of current practice. *Digital Scholarship in the Humanities*, 38(2):530–543.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Zahra Iman, Scott Sanner, Mohamed Reda Bouadjenek, and Lexing Xie. 2017. A longitudinal study of topic classification on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 552–555.

Akanksha Joshi, Eduardo Fidalgo, Enrique Alegre, and Laura Fernández-Robles. 2023. Deepsumm: Exploiting topic models and sequence to sequence networks for extractive text summarization. *Expert Systems with Applications*, 211:118442.

[10] https://www.veraai.eu/home

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Machine learning proceedings 1995*, pages 331–339. Elsevier.

Jey Han Lau, David Newman, Sarvnaz Karimi, and Timothy Baldwin. 2010. Best topic word selection for topic labelling. In *Coling 2010: Posters*, pages 605–613.

Minsuk Lee, Weiqing Wang, and Hong Yu. 2006. Exploring supervised and unsupervised methods to detect topics in biomedical text. *BMC bioinformatics*, 7:1–11.

Tak Yeon Lee, Alison Smith, Kevin Seppi, Niklas Elmqvist, Jordan Boyd-Graber, and Leah Findlater. 2017. The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies*, 105:28–42.

Dai Li, Bolun Zhang, and Yimang Zhou. 2023. Can large language models (llm) label topics from a topic model? *SocArXiv: 10.31235/osf.io/23x4m*.

Yida Mu, Mali Jin, Kalina Bontcheva, and Xingyi Song. 2023a. Examining temporalities on stance detection towards covid-19 vaccination. *arXiv preprint arXiv:2304.04806*.

Yida Mu, Mali Jin, Charlie Grimshaw, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2023b. Vaxxhesitancy: A dataset for studying hesitancy towards covid-19 vaccination on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1052–1062.

Yida Mu, Ben P Wu, William Thorne, Ambrose Robinson, Nikolaos Aletras, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2023c. Navigating prompt complexity for zero-shot classification: A study of large language models in computational social science. *arXiv preprint arXiv:2305.14310*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. Yarn: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*.

Soham Poddar, Azlaan Mustafa Samad, Rajdeep Mukherjee, Niloy Ganguly, and Saptarshi Ghosh. 2022. Caves: A dataset to facilitate explainable classification and summarization of concerns towards covid vaccines. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3154–3164.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*.

Daniel Ramage, Evan Rosen, Jason Chuang, Christopher D Manning, and Daniel A McFarland. 2009. Topic modeling for the social sciences. In *NIPS 2009 workshop on applications for topic models: text and beyond*, volume 5, pages 1–4.

Emil Rijcken, Floortje Scheepers, Kalliopi Zervanou, Marco Spruit, Pablo Mosteiro, and Uzay Kaymak. 2023. Towards interpreting topic models with chatgpt. In *The 20th World Congress of the International Fuzzy Systems Association*.

Alexandra Schofield and David Mimno. 2016. Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics*, 4:287–300.

Chenhui Shen, Liying Cheng, Yang You, and Lidong Bing. 2023. Are large language models good evaluators for abstractive summarization? *arXiv preprint arXiv:2305.13091*.

Xingyi Song, Johann Petrak, Ye Jiang, Iknoor Singh, Diana Maynard, and Kalina Bontcheva. 2021. Classification aware neural topic model for covid-19 disinformation categorisation. *PloS one*, 16(2):e0247086.

Ridam Srivastava, Prabhav Singh, KPS Rana, and Vineet Kumar. 2022. A topic modeled unsupervised approach to single document extractive text summarization. *Knowledge-Based Systems*, 246:108636.

Dominik Stammbach, Vilém Zouhar, Alexander Hoyle, Mrinmaya Sachan, and Elliott Ash. 2023. Re-visiting automated topic model evaluation with large language models. *arXiv preprint arXiv:2305.12152*.

Ihor Stepanov, Arsentii Ivasiuk, Oleksandr Yavorskyi, and Alina Frolova. 2023. Comparative analysis of classification techniques for topic-based biomedical literature categorisation. *Frontiers in Genetics*, 14:1238140.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, et al. 2023. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6(1):158.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ike Vayansky and Sathish AP Kumar. 2020. A review of topic modeling methods. *Information Systems*, 94:101582.

Chong Wang, David Blei, and David Heckerman. 2008. Continuous time dynamic topic models. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, pages 579–586.

Sida I Wang and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 90–94.

Xinyi Wang, Wanrong Zhu, and William Yang Wang. 2023. Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning. *arXiv preprint arXiv:2301.11916*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Ning Wu, Ming Gong, Linjun Shou, Shining Liang, and Daxin Jiang. 2023. Large language models are diverse role-players for summarization evaluation. *arXiv preprint arXiv:2303.15078*.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*.