

# Do We Need Subject Matter Experts? A Case Study of Measuring Up GPT-4 Against Scholars in Topic Evaluation

Kabir Manandhar Shrestha<sup>1,†</sup> (Melbourne Data Analytics Platform), Katie Wood<sup>2,†</sup> (University of Melbourne Archives), David Goodman<sup>3,†</sup> (Historical and Philosophical Studies) and Meladel Mistica<sup>4,†</sup> (Melbourne Data Analytics Platform)

<sup>1</sup>University of Melbourne, Parkville VIC 3010, Melbourne, Australia

## Abstract

Assessing the quality of topics extracted from large text datasets presents a significant challenge in the field of computational social science. This research examines the effectiveness of coherence metrics, the GPT-4 model, and evaluations by subject matter experts (SMEs) using a set of speeches by former Australian Prime Minister Malcolm Fraser. Our primary objective was to analyze the evolution of Fraser's rhetoric. By comparing topics identified by coherence metrics and GPT-4 to those deemed *meaningful* by SMEs, we found that GPT-4 not only performs on par with traditional coherence metrics but also offers a scalable alternative for comprehensive topic evaluations. However, SMEs provide unparalleled depth and contextual understanding, proving indispensable in situations demanding meticulous accuracy. In situations where SMEs aren't available, our approach does show that GPT-4 can be employed for topic evaluation, albeit with some margin of error.

## Keywords

Computational Social Science, Topic Modeling, Latent Dirichlet Allocation (LDA), Coherence Metrics, Large Language Models (LLMs), GPT-4, Evaluation Metrics

## 1. Introduction

The digital era, marked by the fusion of vast data sets and advanced computational techniques, has instigated significant shifts in social and cultural research, ushering in the possibility of novel approaches to studying human behavior and communication [2, 3]. Historically, within Humanities and Social Sciences, *close reading* has been highly valued. This method focuses on small passages, emphasizing lexical choice and structure, as well of course as analyzing entire texts. However, with the integration of computational methods in Social Sciences, the term *distant reading* emerged [4], emphasizing a holistic view of texts or document collections. Topic

---

NL4AI 2023: Seventh Workshop on Natural Language for Artificial Intelligence, November 6-7th, 2023, Rome, Italy [1]

✉ k.manandharshrestha@unimelb.edu.au (K. M. Shrestha); kathrynw@unimelb.edu.au (K. Wood);

d.goodman@unimelb.edu.au (D. Goodman); misticam@unimelb.edu.au (M. Mistica)

🌐 <https://github.com/kabirmanandharsth> (K. M. Shrestha);


<https://findanexpert.unimelb.edu.au/profile/12624-david-goodman> (D. Goodman);

<https://findanexpert.unimelb.edu.au/profile/3575-mel-mistica> (M. Mistica)

🆔 0009-0001-2059-1683 (K. M. Shrestha); 0000-0003-3646-036X (K. Wood)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

modeling, in particular, has gained prominence as a method for this kind of analysis [5, 6]. The adoption of these computational techniques has enabled humanists and social scientists to analyze data on a much broader scale, offering the exciting possibility of entirely fresh perspectives on human behavior and communication [7].

Central to this transformation is computational social science, which harnesses these techniques to decode intricate patterns and trends previously elusive [2]. Within this domain, topic modeling stands out as an adept tool for uncovering latent themes in extensive text corpora, shedding light on cultural evolution, thematic dynamics over time, and its broader applications in computational social science [3, 8, 9].

Historically, topic evaluations have predominantly relied on coherence metrics, emphasizing coherence, relevance, and interpretability. However, as the discipline evolves, there's an increasing awareness that these metrics may not capture the full depth or importance of certain topics [10] and that we need Subject Matter Experts (SMEs) for precision. While SMEs provide invaluable depth, their engagement can be challenging due to availability and expense. In light of these challenges, we explored the potential of GPT-4, a large language model, to emulate the discernment of SMEs in topic evaluations as well.

Our research is centered on the speeches of former Australian Prime Minister Malcolm Fraser, with dual objectives: tracing the trajectory of Fraser's rhetoric and providing a comprehensive view of his political odyssey<sup>1</sup>. To this end, we employed topic models and evaluated their quality through three distinct lenses: intrinsic coherence metrics, insights from subject matter experts (SMEs), and the computational prowess of GPT-4. While this paper focuses on topic evaluation and comparing coherence metrics and GPT-4 with SMEs, readers interested in a deeper dive into our broader research question can refer to Section 8 for our dedicated website and additional resources.

Exploring the novel application of GPT-4 in topic evaluation, we examine its capabilities in assessing topics. We compare GPT-4's assessments with traditional measures of coherence and judgments made by SMEs. Although GPT-4 emerges as a scalable alternative that mirrors the effectiveness of coherence metrics, it occasionally misses subtle details that SMEs recognize. Conversely, the automated analyses may pick up latent patterns of which the SME's were unaware. Based on these findings, we advocate for a collaborative approach to topic evaluation, emphasizing the synergy of human expertise and automated insights. This approach provides researchers with a roadmap for incorporating various evaluation techniques tailored to their specific challenges.

## 2. Background

Malcolm Fraser, the former Prime Minister of Australia, served from 1975 to 1983. Although he was leader of the conservative Liberal Party of Australia, upon leaving office Fraser became estranged from the party and highly critical of its policies and rhetoric., Fraser became a prominent commentator on a range of issues, such as human rights, multiculturalism in Australia, Indigenous affairs, and foreign policy.

To make Fraser's long and varied political life more accessible and digestible, this study employed topic modeling to analyze his radio speeches. The radio speeches (given in english)

---

<sup>1</sup><https://library.unimelb.edu.au/asc/collections/highlights/collections/malcolmfraser>

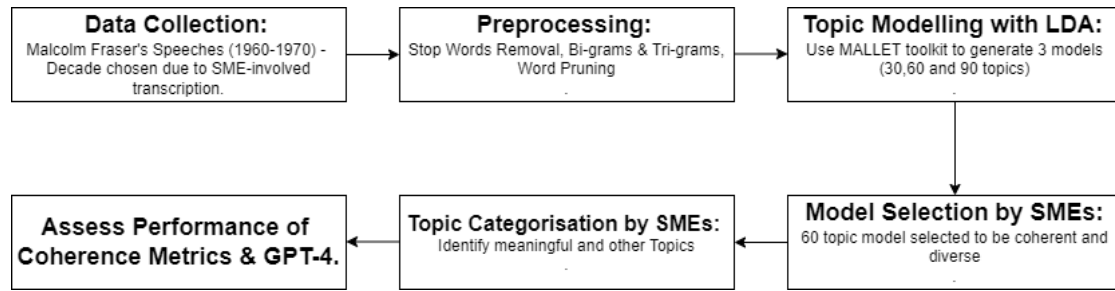
are a unique resource, in that they span his entire parliamentary career, from 1954 to 1983, are relatively consistent in format and wide-ranging in subject matter. The team included the **curator of the Malcolm Fraser Collection**, and an **historian**, who assessed the utility of the induced topics solely based on their expert judgment. Topic modeling, rooted in text mining, has been instrumental across fields from computational social science to digital humanities, enabling scholars to identify latent patterns in vast textual data [11].

Central to the realm of topic modeling is the Latent Dirichlet Allocation (LDA) algorithm [12]. LDA is a probabilistic model that operates on the assumption that each document consists of a combination of different topics, and conversely, each topic is made up of various words. Through careful analysis of how words are distributed across documents, LDA attempts to uncover the underlying topics that could have generated these observed documents [11]. In the digital age, with the rapid growth of vast document archives, the utility of LDA and other topic models becomes even more pronounced, offering algorithmic solutions to manage, organize, and annotate large text collections [13]. LDA's adaptability and robustness have cemented its position as a cornerstone in topic modeling, proving its mettle across diverse datasets and research contexts [14, 15].

Within computational disciplines, topic coherence scores are the accepted measure of topic quality and semantic interpretability with the most common quantitative intrinsic evaluations of topics being pointwise mutual information (PMI) [16] and normalised pointwise mutual information (NPMI) [17]. There are many variations to measure topic coherence, for example Mimno et al. [18] employs log conditional probability (LCP) rather than the accepted (N)PMI, Röder et al. [19] proposes a measure, CV, that combines an indirect cosine measure with NMPI, and Aletras and Stevenson [20] experiments with representing the topic words as context vectors instead of their citation or surface form. Despite the variations, these measures commonly rely on how closely associated the top n words are with each other according to a general reference corpus [17]. In addition, rather than focusing on a specific topic, where a scholar or scholars may have in-depth insights, like our Malcolm Fraser project, in determining measures of coherence and topic quality, topics are often judged by crowd sourced participants [20, 17, 21]. However, in our instance, we have SMEs in this field who would be more equipped in judging topics rather than crowd-sourced workers. We aim to investigate the impact of incorporating SMEs in the topic modeling process by comparing the results of their evaluations to those produced by common quantitative measure.

The last decade has witnessed a paradigm shift in the realm of natural language processing (NLP). This evolution in language understanding can be traced back to the Turing Test's inception in the 1950s, with the journey transitioning from statistical models to the modern pre-trained models that harness the Transformer architecture [22]. These models, characterized by their massive number of parameters and extensive training data, have set new benchmarks across several NLP tasks [23]. As these models scaled, they began to exhibit unique abilities, such as in-context learning, leading to the term "large language models (LLM)" [24]. The research surge in LLMs has reinvigorated discussions on the potential of artificial general intelligence (AGI) [24], emphasizing the blend of human expertise and computational prowess in language modeling [25, 26].

The GPT (Generative Pre-trained Transformer) series [27], developed by OpenAI, stands as a testament to the rapid advancements in this domain. GPT-4 [28], the latest in this lineage, is



**Figure 1:** Topic Modeling Framework

trained on extensive datasets, enabling it to produce human-like text with remarkable accuracy. Its transformer-based architecture excels in capturing textual nuances and context. Beyond just text generation, the potential applications of GPT-4 are vast. From serving as virtual assistants, aiding in content creation, to more specialized tasks like medical diagnosis assistance and legal document analysis, GPT-4's prowess has been demonstrated across domains [29]. For our research purpose, we assess if GPT-4 can act as a subject matter expert. While doing so, we also assess if it fails to align with Subject Matter Experts' point of view and if it provides some misleading information in our case-study as it has in the past [23].

### 3. Topic Analysis Framework

#### 3.1. Data and Objectives

Our research aimed to capture the evolving political landscape in Australia by analyzing the topics addressed by Malcolm Fraser over time. All speeches were delivered in English, reflecting the primary language spoken by the Prime Minister of Australia. We focused on the speeches from the decade spanning 1960 to 1970 because of the relatively low error rate in transcribed documents during these years. This is because Subject Matter Experts (SMEs) were involved in the transcription process.

#### 3.2. Preprocessing for Quality Enhancement

The 270 speeches analyzed comprised a total of 245,701 tokens. Prior to constructing the topic models, we executed several preprocessing steps to enhance the quality of the results. Stop words were eliminated to prevent them from influencing the topic formation process. Additionally, we identified relevant bi-grams and tri-grams to preserve essential word combinations. Moreover, we pruned words and phrases that appeared in less than 10 documents or were present in over 50% of the documents. These measures collectively aimed to enhance the representativeness and clarity of the identified topics. After preprocessing, the total vocabulary size was 1425.

**Table 1**

Examples of 'meaningful' and 'other' topics with subject labels on the left and topic words on the right

Meaningful	Wool Marketing	wool; industry; board; grower; woolgrower; marketing; promotion; conference; levy
	Vietnam War	operation; enemy; viet cong; battalion; regiment; province; village; task force; troop
	Education	school; education; university; technical; student; assistance; high; science; study; training
Other	Unsure/Unknown	concern; affect; problem; important; involve; opportunity; importance; individual; return
	Parliament	operation; parliament; house; minister; day; party; question; speaker; business; sit
	Industry Policy	scheme; argument; price; buy; fact; show; put; fund; market; plan

### 3.3. Topic Modelling Technique

We employed Latent Dirichlet Allocation (LDA) for our topic modeling needs. LDA's clear interpretability, scalability, and broad acceptance in academic research made it a fitting choice for analyzing Malcolm Fraser's speeches. For our analysis, we utilized the MALLET toolkit<sup>2</sup>, a prominent implementation of LDA, to generate three topic models: 30, 60, and 90 topics in a grid search approach in our 270 speeches. We chose three models to provide a range of granularity in the topics, allowing for a more comprehensive analysis of the corresponding topics discussed by Fraser over time.

### 3.4. Model Selection

Selecting the optimal topic model was a crucial step in our methodology. To ensure an informed decision, we enlisted the expertise of SMEs to assess the three generated models. The evaluation process involved analyzing the coherence and diversity of each topic model. Ultimately, based on their assessments, the SMEs concluded that the model with 60 topics best suited the content of Fraser's speeches, striking a balance between granularity and comprehensiveness.

### 3.5. Expert-Based Topic Categorization: Foundation for Comparative Analysis

Our journey into comparative analysis was underpinned by a pivotal phase where the expertise of SMEs came to the forefront. With the aid of the 60-topic model chosen by our SMEs, each topic was carefully categorized, setting the stage for assessing different evaluation methods. These SMEs, equipped with a profound understanding of Malcolm Fraser's political journey, worked in cohesion to separate *meaningful* topics from the rest. A *meaningful* topic was one that provided practical, valuable, and coherent insights into various aspects of Fraser's political trajectory.

<sup>2</sup><https://mimno.github.io/Mallet/topics>

n	MEASURE	Precision	Recall	F1-Score
20	PMI	<b>0.85</b>	0.46	0.60
	NPMI	<b>0.85</b>	0.46	0.60
	LCP	0.48	0.51	0.49
30	PMI	<b>0.83</b>	0.68	0.75
	NPMI	<b>0.83</b>	0.68	0.75
	LCP	0.77	0.62	0.69
40	PMI	0.78	<b>0.84</b>	<b>0.81</b>
	NPMI	0.78	<b>0.84</b>	<b>0.81</b>
	LCP	0.48	0.51	0.49
	ORACLE-NPMI	0.81	0.81	0.81
	RANDOM	0.61	0.50	0.55
	MAJORITY	0.62	1.0	0.76

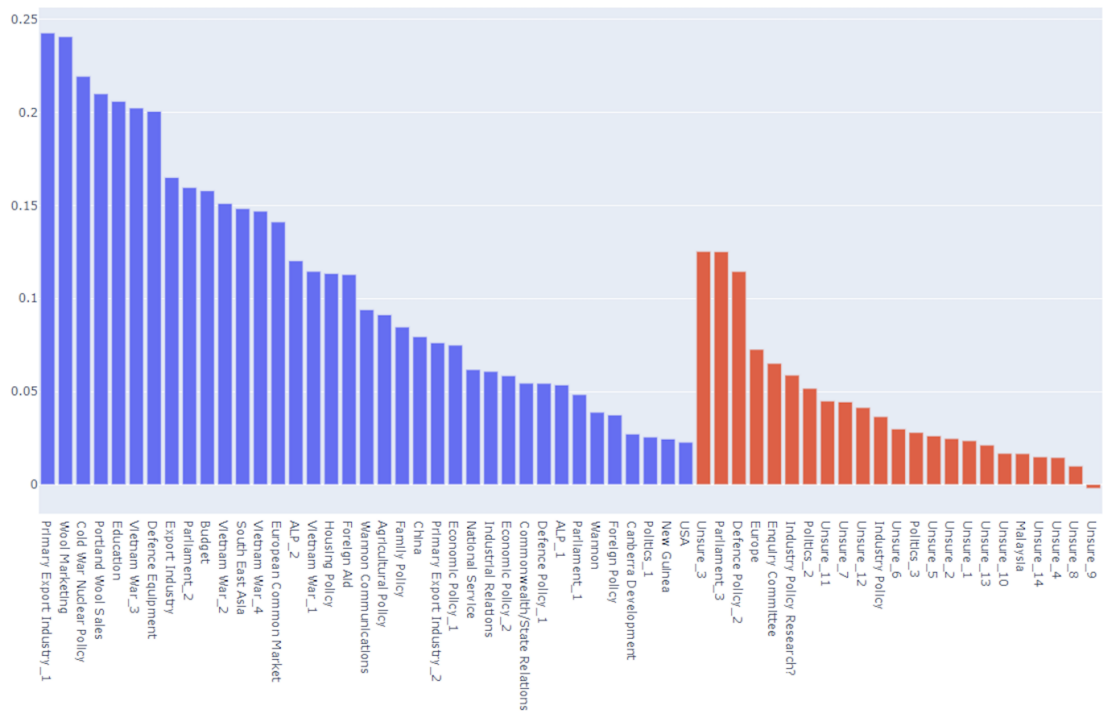
**Table 2**  
Coherence as predictors of *meaningful* topics for the top n-ranked topics

On the other hand, *other* topics encompassed themes that, while coherent, did not seamlessly harmonize with the overarching narrative of Fraser’s political journey. SMEs further enriched our analysis by appending subject labels to topics, where feasible. These labels encompassed diverse subjects, ranging from the Vietnam War and Wool Marketing to Education, reflecting the wide spectrum of themes in Fraser’s speeches. Examples of both *meaningful* and *other* topics can be seen in Table 1.

This classification procedure takes center stage in our study, serving as the foundation against which the performance of both quantitative metrics and GPT-4 is compared. The SMEs’ keen evaluation of topic significance forms a pivotal reference point, enabling us to gauge the accuracy and subtleties of alternative approaches. This strategic alignment underscores the SMEs’ role not merely as annotators but as navigators guiding our exploration of automated evaluation methods. An overall summary of our framework is seen in Figure 1.

#### 4. Comparative Analysis of Coherence Metrics, Subject Matter Experts, and GPT-4: A Multifaceted Evaluation

In this section, we delve into an extensive comparative analysis that assesses the efficacy of coherence metrics, the discernment of subject matter experts (SMEs), and the automated capabilities of GPT-4. Our investigation aims to determine the extent to which these evaluation methodologies can effectively identify the significance of topics derived from our Latent Dirichlet Allocation (LDA) model.



**Figure 2:** NPMI topic coherence: blue bars on left represent *meaningful* topics; red bars on right are *other* topics.

#### 4.1. Coherence Metrics: A Quantitative Lens on Topic Significance

To determine how well the quantitative evaluation of topics measured up against the judgments of our SMEs, we calculated 3 coherence scores: Pointwise Mutual Information (PMI), Normalized Pointwise Mutual Information (NPMI), and Local Coherence Probability (LCP) as calculated in Lau and Baldwin [21]. PMI measures the pointwise mutual information between each pair of words in a topic, while NPMI normalizes this score by dividing it by the logarithm of the probability of the two words occurring together. LCP, on the other hand, measures the probability of a given word following another word within a certain distance in the same document. These coherence scores were used as a predictor of topic quality.

For the reference corpus, we created a corpus that consisted of 2,500 Wikipedia articles, 5,000 news corpus articles, and 484 Malcolm Fraser speeches that were not used in the modeling of the topics nor the decade from which the topics had been induced. The reference corpus serves as a baseline for evaluating the quality of the topics [17] which allowed us to evaluate the coherence of the topics generated by the LDA model against the opinions of subject matter experts (SMEs). The corpus had approximately 99k tokens with 47k, 32k, and 20k being made up of the news corpus, Wikipedia, and the remaining Malcolm Fraser speeches, respectively. Specifically, we wanted to ascertain if high coherence scores correlated with *meaningful* topics, so we ranked them according to their coherence scores (highest to lowest) and labeled the top n-ranked topics



( $n = 20, 30, 40$ ) as *meaningful*. The reason for the selection of these values could be based on a trade-off between the number of topics that can be labeled as *meaningful* and the quality of the labeled topics. The results of coherence as a predictor of *meaningful* topics are shown in Table 2. As an upper bound, we also labeled the top 37 topics of the ranked NPMI, which coincides with the number of topics the SMEs deemed *meaningful*, shown as ORACLE-NPMI. This table shows that the coherence scores for NPMI and PMI consistently performed better than our RANDOM and MAJORITY class baselines, with NPMI achieving an f1-score of 0.81 for  $n=40$ , equivalent to the f1-score for ORACLE-NPMI. The metric LCP did not always surpass these baselines and came nowhere near ORACLE-NPMI. Save for LCP, using quantitative topic coherence measures to ascertain quality topics would have resulted in comparable topics to the ones chosen by SMEs for  $n=40$ .

Figure 2 shows the coherence score for the NPMI metric with the blue topics, on the left, deemed *meaningful*, and the red topics were those marked as *other*. As can be seen in this graphic, the *meaningful* topics consistently have higher coherence scores, which coincides with them being good predictors of *meaningful* topics as deemed by SMEs. We performed the Kruskal-Wallis test [30] to determine whether there is an inherent difference in the *meaningful* and *other* topics seen in Figure 2. We calculated  $H = 17.93$  and  $p\text{-value} = 2.29e^{-5}$  using the scipy Python library. Therefore, given the small p-value, we accept the null hypothesis that the distribution of the two groups is not the same.

## 4.2. GPT-4 as a Surrogate SME: Automating Topic Classification

The evaluation of topic models often hinges on the coherence and relevance of the topics generated. Traditional metrics like NPMI and PMI offer quantitative measures, but as our findings suggest, they occasionally fall short in capturing the nuanced understanding that Subject Matter Experts (SMEs) bring to the table. For instance, topics such as ‘USA’ and ‘Foreign Policy’, despite their apparent significance, might be overlooked based on these metrics alone, as seen in Figure 2.

The real-world application of topic modeling, especially in specialized domains, often requires the insights of SMEs. Their deep domain knowledge can reveal subtleties and intricacies that might elude quantitative metrics. However, the engagement of SMEs is not without its challenges. Their expertise, though invaluable, might not always be readily available. Moreover, the evaluation process can be labor-intensive and time-consuming. A pivotal challenge is determining the optimal value for  $n$ , the number of top-ranked topics labeled as *meaningful*. This choice is inherently subjective and can significantly influence the outcome of the evaluation, potentially introducing variability and inconsistencies in the insights derived.

To circumvent these challenges and introduce a scalable, automated approach to topic classification, we turned to the capabilities of large language models, specifically GPT-4. The underlying hypothesis was simple yet profound: Could a model like GPT-4, with its vast training data and sophisticated architecture, emulate the discernment typically associated with SMEs?

Our methodology involves tasking GPT-4 with the classification of the 60 identified topics. As prompts, we provided the model with a detailed context of our collaboration, explaining the aim, the modeling process, and the SMEs’ evaluation criteria. The model was then presented with the top 10 words for each of the 60 topics just as SMEs’ were and was tasked with classifying the topics as *meaningful* or *other*. To ensure the robustness of our approach and to account



Temperature	Precision	Recall	F1-Score
0	0.73	<b>0.89</b>	0.80
0.25	0.73	<b>0.89</b>	0.80
0.5	<b>0.76</b>	0.86	<b>0.81</b>
0.75	0.72	<b>0.89</b>	0.80
1	0.74	0.86	0.80
ORACLE-NPMI	0.81	0.81	0.81
RANDOM	0.61	0.50	0.55
MAJORITY	0.62	1.0	0.76

**Table 3**

Comparative analysis of ORACLE-NPMI and GPT-4 at different temperature settings in classifying *meaningful* and *other* topics.

for potential model variability, each topic undergoes classification five times. This iterative methodology not only provides multiple perspectives on each topic but also allows us to assess the internal consistency of GPT-4’s classifications. We chose the maximum result from five runs.

Recognizing the potential variability in GPT-4’s outputs, we explored different temperature settings. The temperature parameter in GPT-4 influences the randomness of the model’s outputs. A lower temperature makes the model’s predictions more deterministic, while a higher temperature introduces more variability. By experimenting with a range of temperature values, from 0 to 1, we aim to identify the optimal setting that maximizes the model’s accuracy while preserving its ability to discern *meaningful* topics. Notably, while GPT-4 offers two primary parameters to influence its output randomness - temperature and top\_p - we chose to adjust only the temperature setting. This decision was based on the GPT-4 API documentation from OpenAI, which recommends adjusting either the temperature or top\_p, but not both<sup>3</sup>.

Table 3 offers a comparative analysis, stating the performance metrics of GPT-4 against ORACLE-NPMI for the classification of *meaningful* and *other* topics. This comparison underscores the potential of GPT-4 as a surrogate SME. The results indicate that GPT-4, at a temperature of 0.5, can match the F Score of the Oracle NPMI with a score of 0.81, and achieves an F score of 0.8 at all other temperatures. Importantly, while overall accuracy is crucial, our primary focus was on correctly identifying as many *meaningful* topics as possible, given their significance in describing Fraser’s narrative. Notably, at all temperatures, GPT-4 surpassed the ORACLE-NPMI’s recall. The ability to achieve this performance without prior knowledge of the optimal value for **n** underscores the promise of this approach.

<sup>3</sup><https://platform.openai.com/docs/api-reference/chat/create>

## 5. Discussion

Only SMEs	Canberra Development	canberra; capital; decision; move; development; building; years ago; grow; hear; home
	Politics	programme; point; view; remark; speak; line; attack; question; full; press
	Parliament	opposition; debate; majority; attack; week; thing; question; show; move; important
SMEs and GPT-4	Wannon	fraser; company; operate; malcolm fraser; today; worth; wannon; order; support; press
	USA	united states; american; president; interest; world; thing; washington; meet
	Foreign Policy	prime minister; leader; world; international; difference; president; united nation; whitlam; asian; put
	New Guinea	new guinea; agreement; full; council; position; territory; white; bring; long; argument
SMEs and ORACLE-NPMI	National Service	call; service; period; provide; introduce; full; condition; general; provision; additional
	ALP	cost; speech; calwell; general; thing; kind; promise; hard; governor; pay

**Table 4**

Comparison of subset of topics identified as *meaningful* by different methods : Subject Matter Experts(SMEs), ORACLE-NPMI and GPT-4

The evaluation of topic models, especially in the context of historical speeches like those of Malcolm Fraser, presents a unique set of challenges and considerations. Our comparative analysis between NPMI, GPT-4, and SMEs has shed light on the intricacies of topic classification, revealing both the strengths and limitations of each approach. Our findings as highlighted in Table 4 provide a comprehensive understanding of the differences between the approaches. In the table, we specifically showcase the results from GPT-4 at a temperature value of 0.5, which achieved the highest F1-score of 0.81 while also surpassing the recall of ORACLE-NPMI at 0.86. Let's delve deeper:

- **Performance Metrics and Their Implications:**

1. **NPMI's Robustness:** NPMI's strong performance in terms of precision, recall, and F1-score reinforces its reputation as a dependable quantitative coherence measure. Its alignment with SMEs on topics like "National Service" and "ALP" underscores its value in topic modeling tasks and its ability to capture significant political and policy narratives. A key factor contributing to NPMI's success in identifying these topics is the reference corpus used. A well-curated reference corpus enhances the accuracy and relevance of coherence-based topic modeling by capturing context.
  2. **GPT-4's Competence:** The results showing GPT-4's performance on par with NPMI is both remarkable and promising. Its alignment with SMEs on topics such as "Wannon" and "Foreign Policy" indicates the vast potential of large language models in domain-specific tasks. However, while GPT-4 can emulate the capabilities of specialized metrics, it may not possess the same depth of understanding as SMEs, as evident in its unique topic identifications.
  3. **Patterns in GPT-4's Misclassifications:**
    - a) **Context-Specific Topics:** Topics such as "Canberra Development," "National Service," and "ALP" as seen in Table 4 are deeply rooted in the Australian context. GPT-4 might have struggled with these due to their specificity and the potential lack of emphasis on the Australian political and developmental context in its training data.
    - b) **Broad and Abstract Topics:** "Politics" and "Parliament" are more abstract and encompassing. While SMEs can easily discern the nuances of such topics, GPT-4 might have found it challenging to pinpoint their significance amidst other potential topics.
- **Scalability, Consistency, and Real-world Applicability:** In practical scenarios, handling large volumes of data demands scalable solutions. GPT-4's capacity to efficiently process extensive datasets without sacrificing consistency makes it an invaluable tool for large-scale text analysis tasks. This scalability becomes even more essential when obtaining SME insights is challenging due to various constraints. GPT-4's consistent outputs are commendable. However, the occasional variations we observed in topic labeling highlight the importance of iterative methodologies and cross-validation in topic classification endeavors.
  - **The Indispensable Role of SMEs:** The unique topics identified by SMEs, such as "Canberra Development" and "Politics," emphasize their unparalleled depth of understanding. Beyond mere numbers, SMEs contribute qualitative insights, adding layers of interpretation and context that enrich the overall analysis. Their expertise becomes indispensable when dissecting content like Fraser's speeches, where historical, political, and cultural nuances significantly influence the overarching narrative.
  - **Future Directions and Considerations:** One of the most promising avenues for future research lies in the synergy between SMEs and models like GPT-4. SMEs bring unparalleled depth and understanding, especially when determining the granularity of topics needed to address broader research questions, such as discerning shifts in Fraser's rhetoric over time. Once this granularity is established, leveraging GPT-4 can be invaluable. The

model's scalability and consistency can be employed to assess the significance of these topics, streamlining the process and reducing the manual effort required by SMEs albeit with some error.

## 6. Limitations

A notable limitation of our study is its reliance on Subject Matter Experts (SMEs) possessing profound knowledge of the dataset. Individuals lacking this specialized expertise might encounter challenges in selecting the optimal number of topics or in evaluating the clarity and significance of each topic. The Latent Dirichlet Allocation (LDA) method we employed conceptualizes each document as a blend of topics, presuming each word originates from one of these topics. However, this assumption might not hold true universally across diverse document types or topics, potentially influencing the quality of the topics identified. Exploring alternative methodologies, such as BERTopic [31] and Topic2Vec [32], could provide valuable comparative insights. While we did not venture into these methods within the scope of this study, they present promising avenues for future research.

Additionally, our analysis is temporally constrained, focusing on the period from 1960 to 1970 due to the limited availability of Fraser's speeches. Extending the analysis to encompass speeches from varied time frames could validate the generalizability of our findings.

Another dimension worth highlighting is our exclusive utilization of the GPT-4 Large Language Model (LLM). We did experiment with other models like Llama 2 chat and PALM 2, but encountered challenges. Specifically, these models struggled to classify topics using the prompts that were effective for GPT-4. The prompt length was also a limiting factor, which was not an issue with GPT-4. Crafting uniform prompts compatible with several Large Language Models would have necessitated significant time and effort. It's essential to clarify that the primary objective of this paper was not a comparative analysis of multiple LLMs, but rather an exploration of the feasibility of using LLMs as potential replacements for traditional coherence metrics. The results derived from GPT-4 affirmatively indicate this possibility.

## 7. Conclusion

In this study, we embarked on a journey to understand the comparative efficacy of subject matter experts (SMEs), coherence metrics, and the capabilities of GPT-4 in evaluating topic models. While our initial expectation leaned towards SMEs being the gold standard for topic evaluation in a specific domain, our findings revealed that common measures like coherence metrics performed commendably, often aligning closely with expert judgments. Nevertheless, relying solely on these metrics could have led to overlooking some pivotal topics. The introduction of GPT-4 into the evaluation mix not only showcased its potential as a scalable and consistent evaluator but also highlighted its ability to match, and in some instances, rival traditional coherence metrics. Yet, even with GPT-4's impressive performance, the study underscores the irreplaceable value of SMEs' nuanced understanding. In essence, while tools like coherence metrics and GPT-4 offer promising avenues for topic evaluation, they should complement, not replace, the insights of SMEs. Our study advocates for a balanced amalgamation of automated methods and human

expertise for comprehensive topic model evaluation, emphasizing that GPT-4 can serve as a standalone evaluator in the absence of SMEs.

## References

- [1] E. Bassignana, D. Brunato, M. Polignano, A. Ramponi, Preface to the Seventh Workshop on Natural Language for Artificial Intelligence (NL4AI), in: Proceedings of the Seventh Workshop on Natural Language for Artificial Intelligence (NL4AI 2023) co-located with 22th International Conference of the Italian Association for Artificial Intelligence (AI\* IA 2023), 2023.
- [2] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, Life in the network: The coming age of computational social science 323 (2009).
- [3] Y. Liu, X. Feng, Y. Zhang, Y. Kong, R. Yang, Paths study on knowledge convergence and development in computational social science: Data metric analysis based on web of science, *Complexity* 2022 (2022) 1–18. doi:10.1155/2022/3200371.
- [4] F. Moretti, Conjectures on world literature, *New Left Review* 1 (2000) 54–68.
- [5] J. Taylor, B. Adams, The wild process: constructing multi-scalar environmental narratives, Ubiquity Press Ltd, United Kingdom, 2022.
- [6] J. Taylor, M. Mistica, G. Fairclough, T. Baldwin, Inferring Value: A Multiscalar Analysis of Landscape Character Assessments, Ubiquity Press Ltd, United Kingdom, 2022.
- [7] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, M. Van Alstyne, Computational social science, *Science* 323 (2009) 721–723.
- [8] M. Peponakis, S. Kapidakis, M. Doerr, E. Tountasaki, From calculations to reasoning: History, trends and the potential of computational ethnography and computational social anthropology, *Social Science Computer Review* (2023) 089443932311676. doi:10.1177/08944393231167692.
- [9] J. Boyd-Graber, Y. Hu, D. Mimno, Applications of topic models, *Foundations and Trends® in Information Retrieval* 11 (2017) 143–296. doi:10.1561/15000000030.
- [10] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, D. Blei, Reading tea leaves: How humans interpret topic models, *Neural Information Processing Systems* 22 (2009) 288–296.
- [11] H. Jelodar, Y. Wang, C. Yuan, X. Feng, Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey (2017).
- [12] D. M. Blei, A. Ng, M. I. Jordan, Latent dirichlet allocation, *J. Mach. Learn. Res.* 3 (2001) 993–1022.
- [13] D. M. Blei, Probabilistic topic models, *Communications of the ACM* 55 (2012) 77 – 84. URL: <https://api.semanticscholar.org/CorpusID:753304>.
- [14] R. Alghamdi, K. Alfalqi, A survey of topic modeling in text mining, *International Journal of Advanced Computer Science and Applications* 6 (2015). doi:10.14569/IJACSA.2015.060121.
- [15] T. Silwattanusarn, P. Kulkanjanapiban, A text mining and topic modeling based bibliomet-

- ric exploration of information science research, *IAES International Journal of Artificial Intelligence (IJ-AI)* 11 (2022) 1057–1065. doi:10.11591/ijai.v11.i3.pp1057-1065.
- [16] D. Newman, J. H. Lau, K. Grieser, T. Baldwin, Automatic evaluation of topic coherence, in: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Los Angeles, California, 2010, pp. 100–108. URL: <https://aclanthology.org/N10-1012>.
- [17] J. H. Lau, D. Newman, T. Baldwin, Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality, in: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Gothenburg, Sweden, 2014, pp. 530–539. URL: <https://aclanthology.org/E14-1056>. doi:10.3115/v1/E14-1056.
- [18] D. Mimno, H. Wallach, E. Talley, M. Leenders, A. McCallum, Optimizing semantic coherence in topic models, in: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Edinburgh, Scotland, UK., 2011, pp. 262–272. URL: <https://aclanthology.org/D11-1024>.
- [19] M. Röder, A. Both, A. Hinneburg, Exploring the space of topic coherence measures, *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (2015)*.
- [20] N. Aletras, M. Stevenson, Evaluating topic coherence using distributional semantics, in: *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, Association for Computational Linguistics, Potsdam, Germany, 2013, pp. 13–22. URL: <https://aclanthology.org/W13-0102>.
- [21] J. H. Lau, T. Baldwin, The sensitivity of topic coherence evaluation to topic cardinality, in: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, San Diego, California, 2016, pp. 483–487. URL: <https://aclanthology.org/N16-1057>. doi:10.18653/v1/N16-1057.
- [22] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: *NIPS, 2017*. URL: <https://api.semanticscholar.org/CorpusID:13756489>.
- [23] A. Thirunavukarasu, Large language models will not replace healthcare professionals: curbing popular fears and hype, *Journal of the Royal Society of Medicine* 116 (2023) 1410768231173123. doi:10.1177/01410768231173123.
- [24] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J. Nie, J. rong Wen, A survey of large language models, *ArXiv abs/2303.18223 (2023)*. URL: <https://api.semanticscholar.org/CorpusID:257900969>.
- [25] L. Floridi, M. Chiriatti, Gpt-3: Its nature, scope, limits, and consequences, *Minds and Machines* 30 (2020) 681–694. URL: <https://api.semanticscholar.org/CorpusID:228954221>.
- [26] K. Ethayarajh, How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings, in: *Conference on Empirical Methods in Natural Language Processing*, 2019. URL: <https://api.semanticscholar.org/CorpusID:202120592>.

- [27] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training (2018).
- [28] OpenAI, Gpt-4 technical report, ArXiv abs/2303.08774 (2023).
- [29] N. Jethani, S. Jones, N. Genes, V. Major, I. Jaffe, A. Cardillo, N. Heilenbach, N. Ali, L. Bonanni, A. Clayburn, Z. Khera, E. Sadler, J. Prasad, J. Schlacter, K. Liu, B. Silva, S. Montgomery, E. Kim, J. Lester, N. Razavian, Evaluating chatgpt in information extraction: A case study of extracting cognitive exam dates and scores (2023). doi:10.1101/2023.07.10.23292373.
- [30] W. H. Kruskal, W. A. Wallis, Use of ranks in one-criterion variance analysis, *Journal of the American Statistical Association* 47 (1952) 583–621.
- [31] M. Grootendorst, Bertopic: Neural topic modeling with a class-based tf-idf procedure, 2022. arXiv:2203.05794.
- [32] L.-Q. Niu, X.-Y. Dai, Topic2vec: Learning distributed representations of topics, 2015. arXiv:1506.08422.

## 8. Online Resources

- Website of our collaboration - Political Rhetoric Project
- Malcolm Fraser's Speeches