Can LLMs categorize the specialized documents from web archives in a better way?

Saran Pandian Pandi
spand43@uic.edu
University of Illinois, Chicago
Chicago, Illinois, USA

Seoyeon Park
seoyeonpark@hanyang.ac.kr
Hanyang University (ERICA)
Ansan, Republic of Korea

Praneeth Rikka
PraneethRikka@my.unt.edu
University of North Texas
Denton, Texas, USA

Cornelia Caragea
cornelia@uic.edu
University of Illinois, Chicago
Chicago, Illinois, USA

Mark E. Phillips
Mark.Phillips@unt.edu
University of North Texas
Denton, Texas, USA

ABSTRACT

The explosive growth of web archives presents a significant challenge: manually curating specialized document collections from this vast data. Existing approaches rely on supervised techniques, but recent advancements in Large Language Models (LLMs) offer new possibilities for automating collection creation. Large Language Models (LLMs) are demonstrating impressive performance on various tasks even without fine-tuning. This paper investigates the effectiveness of prompt design in achieving results comparable to fine-tuned models. We explore different prompting techniques for collecting specialized documents from web archives like UNT.edu, Michigan.gov, and Texas.gov. We then analyze the performance of LLMs under various prompt configurations. Our findings highlight the significant impact of incorporating task descriptions within prompts. Additionally, including the document type as justification for the search scope leads to demonstrably better results. This research suggests that well-crafted prompts can unlock the potential of LLMs for specialized tasks, potentially reducing reliance on resource-intensive fine-tuning. This research paves the way for automating specialized collection creation using LLMs and prompt engineering.

CCS CONCEPTS

• **Information systems** → **Digital libraries and archives**; • **Computing methodologies** → **Natural Language Processing**.

KEYWORDS

Large Language Models, Web archiving, specialized collection, K-shot prompting, Chain-of-thoughts prompting

ACM Reference Format:
Saran Pandian Pandi, Seoyeon Park, Praneeth Rikka, Cornelia Caragea, and Mark E. Phillips. 2024. Can LLMs categorize the specialized documents from web archives in a better way?. In *The 2024 ACM/IEEE Joint Conference on Digital Libraries (JCDL '24), December 16–20, 2024, Hong Kong, China.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3677389.3702591

JCDL '24, December 16–20, 2024, Hong Kong, China

ACM ISBN 979-8-4007-1093-3/24/12
https://doi.org/10.1145/3677389.3702591

1 INTRODUCTION

Web archiving is essential for preserving and providing future access to the content made available across the web. Cultural heritage institutions including libraries, archives, museums and galleries have worked for nearly three decades to preserve the born-digital and digitized resources that are shared with the general public using the World Wide Web.

Web archives serve as a mechanism to provide users with access to resources that may no longer be available on the live web. These resources can be websites, multimedia content, social media posts, publications, or documents. While many organizations use web archiving to provide access to the look and feel of archived websites, some look at web archiving as an automated way of collected resources that are have historically been collected at their institutions but which are no longer made available in physical format [23].

The quantity of web-archived documents has seen a rise in the last 3 decades. For example, Jefferson Bailey of the Internet Archive noted that as of 2016, there were 1.6 billion PDF files in the Global Wayback Machine [3]. With the increase in the quantity of documents and publications in web archives, there is a need to automate the process of identifying the documents that belong to a particular collection [12]. For example, the Library of Congress Web Archiving holdings contains 42,188,995 unique PDF documents in its collections [27]. The documents that belong to a particular collection might be as small as 1% of this archive which is 0.4 million.

A practical example of this type of collection building from web archives can be seen in the web archives collected as part of the Texas Records and Information Locator (TRAIL) which was developed to provide access to Texas state government documents and electronic publications that have been made available to the public through the Internet by or on behalf of a state agency [14]. Currently the TRAIL program uses the web archiving services provided by Archive-It [1] to collect and provide access to these archived resources. Over time the publications and government documents in this web archive could be identified and then described with

[1] TRAIL - https://www.archive-it.org/collections/414

item-level bibliographic metadata as part of a digital library collection such as the Texas State Publications collection [2] in The Portal to Texas History [3] operated by the University of North Texas.

To search and annotate the documents that are in-scope of a collection is often tedious, time-consuming, and expensive. In this paper, we have experimented to know how Large Language Models (LLMs) can categorize specialized documents from a collection of documents extracted from web archives. Large Language Models, as a consequence of a large number of parameters, have been shown to exhibit strong language understanding and generation capabilities. In addition to these, Large language models exhibit emergent capabilities, such as in-context learning, allowing them to perform downstream tasks through mere text-based instruction without explicit fine-tuning [5].

Our work focuses on using this in-context learning capability of LLMs to categorize the specialized documents from web archives through different prompting techniques. We hypothesize that meticulously crafted prompts incorporating sufficient contextual information can obviate the necessity for fine-tuning, as traditional fine-tuning methods are data-hungry and demonstrate diminishing performance with increased context length [8]. Specifically, we try to leverage the few shot-prompting and Chain-of-thought prompting [38] techniques that help LLMs with more information on understanding how to better categorize the documents. These methods employ a small set of exemplar documents to illustrate the task to the LLM, enabling it to comprehend the objective and generate appropriate outputs for the subsequent document. Leveraging the emergent capabilities of LLMs has the potential to mitigate the need for extensive datasets, as these models can often adapt to tasks without explicit fine-tuning [5]. To empirically assess this hypothesis, we compare the performance of LLMs utilizing their inherent abilities with supervised learning approaches such as fine-tuning BERT and RoBERTa that rely on relatively larger datasets.

Our contributions are as follows: **1)** Our research extends the work of [28] by introducing a novel web archive dataset compiled from Michigan state publications. This dataset complements the existing collections curated from scholarly repositories like UNT.edu and government websites like Texas.gov. This dataset is made available for further research in this area[4]. **2)** We conduct a comparative analysis of Zero-shot, One-shot, and Chain-of-thought prompting methodologies to determine the most effective approach for eliciting desired outputs from the language model. The performance of these prompting techniques is subsequently compared to a supervised learning baseline. **3)** We investigated the correlation between document length and model performance. **4)** We examine the robustness of One-shot and Chain-of-thought prompting by assessing model performance under varying exemplar sets. **5)** We examine the impact of increasing the number of exemplars on model performance by comparing One-shot, Two-shot, and Five-shot prompting techniques.

[2]Texas State Publications - https://texashistory.unt.edu/explore/collections/TXPUB/
[3]The Portal to Texas History - https://texashistory.unt.edu
[4]https://www.cs.uic.edu/~cornelia/datasets/web$_{archive_data}$

2 RELATED WORK

This section provides an overview of previous research on automated specialized document collection from web archives. Additionally, we examine relevant studies exploring the application of language models and prompting techniques used in LLMs.

Web archiving: The methods and approaches behind web archiving have been studied widely by people in library sciences [20]. Ntoulas et al. [24] observed that more than 80% of the web pages get updated or removed after 1 year which makes it essential to perform web archiving. However, as the quantity of web-archived documents keeps growing [11] it becomes challenging to filter the desired webpages. This filtering process can be automated using text classification techniques. Traditional text classification techniques such as Support Vector Machine [16] and Naïve Bayes Multinomial [21] methods require specific extracted features such as Bag of words (BoW) or *tf-idf* [7, 33] for training a classification model. [4, 13] proposed language modeling with neural networks which reduced the need for feature engineering techniques. Kalchbrenner et al. [17] introduced a deep learning model with multiple convolutional layers that used random vectors for word embeddings. Zhang and Wallace [40] uses a single layer at the beginning of a convolutional neural network model to convert the token to word embeddings. For identifying the scope of a document from a web archive, several papers have tried traditional machine learning and deep learning models for text classification.

Ayala and Caragea [2] uses traditional feature extraction techniques such as BoW and *tf-idf* features and machine-learning techniques such as support vector and naive Bayes model to identify research articles that are related to the topic of Web Archiving from a collection of documents obtained by crawling the Web. Caragea et al. [6] introduces structural features, which help in identifying the document type and scope of the document. Patel et al. and Patel et al. [28, 29] discusses feature extraction with traditional machine learning models and compares it to deep learning techniques with structural features for identifying the documents in the scope of three repositories namely texas.gov, usda.gov, and unt.edu.

Language Models: Vaswani et al. [36] developed a transformer architecture that was used for machine translation tasks. The transformer architecture was used to build complex language models which were used for natural language understanding such as classification tasks and natural language generation such as machine translation. These models were called Transformer based pretrained language models (PLM). PLMs are further divided into 3 types of architecture [22]; Encoder-only models, decoder-only models and encoder-decoder models. Encoder-only models include the Masked language models (MLM) such as BERT [10], RoBERTa [19]. These language models are used for downstream tasks such as classification, question-answering, etc. Decoder-only models include Causal language models (CLMs) such as GPT [30], GPT-2 [31]. These models due to their causal nature are widely used for Natural language generation tasks such as machine translation and paraphrase generation. Encoder-decoder models have an encoder to encode the input representation and a decoder to generate causal text. Models such as T5 [32] and BART [18]. In addition to text-to-text generation tasks, they are used for other tasks such as image captioning, speech-to-text, etc.

Prompts without context

a. Zero-shot prompting without context:

Identify if the following document is in scope of UNT Scholarly PDF repository or not. Answer 'True' if it belongs and 'False' if it does not belong. Answer in just one word 'True' or 'False'

Document : **(Document)**

Answer :

b. One-shot prompting without context:

complete the following prompt by answering 'In scope:' for the given above 'Document:' Answer 'True' if it is in scope of UNT repository else answer 'False' Here are few examples:

Document : ABSTRACT Cost/Benefit Analysis and Ad Valorem Tax Benefits of Oil and Gas Drilling in the DFW Barnett Shale of Urban and Suburban North Texas John S. Baen Ph.D., University of North Texas. Key words/concepts: oil and gas valuation methods, need for education by real estate related professionals, planners and city administrators, city drilling ordinances, land-use efficiencies. This is a draft.

In scope : True

Document : Embedded Systems Design CSCE 3612, Section 001 and 002 Spring 2016 Class Timings: Tuesday and Thursday, 11:30 AM 12:50 PM, NTDP B142 Instructor: Robin Pottathuparambil Office Hours: Tuesday and Thursday 5:00 PM 6:00 PM or by appointment Instructional Assistant: Evan Rodrigues

In scope: False

Document : **(Document)**

In scope :

Prompts with context

c. Zero-shot prompting with context:

"UNT Scholarly Works is a special collection of items contributed by the UNT Community and hosted in the UNT Digital Library. This collection brings together articles, papers, presentations, books, chapters, reviews, academic posters, artwork, and other scholarly and creative works and makes them readily accessible to showcase UNT's research and creative achievements to a worldwide audience. The UNT Scholarly Works collection also serves as the open access repository for UNT."

From the description above, .Identify if the following document is in scope of UNT Scholarly PDF repository or not. Answer 'True' if it belongs and 'False' if it does not belong. Answer in just one word 'True' or 'False'

Document : **(Document)**

Answer :

Figure 1: a. Zero-shot without context b. One-shot without context c. Zero-shot with context

LLMs: Large Language Models (LLMs) are transformer based PLMs with more than tens to hundreds of billions of parameters. LLMs when given optimal prompts have been shown to perform well on all NLP downstream tasks thus reducing the need to rely on the models trained from scratch for specific tasks. LLMs such as ChatGPT [25], Gemini [34] have been commercialized in recent times. These commercial models have been aligned to human preferences by using Reinforcement Learning with Human Feedback [26] that improves the emergent abilities of the model. Causal language models such as GPT and GPT-2 have demonstrated proficiency in language understanding and generation. However, GPT-3 has exhibited novel capabilities through in-context learning, surpassing the capabilities of its predecessors. Brown et al. [5] showed that through in-context learning where the model is capable of classifying documents based on the examples provided in the prompts, GPT-3 performed better than GPT and GPT-2. Moreover, they have

d. One-shot prompting with context:

"UNT Scholarly Works is a special collection of items contributed by the UNT Community and hosted in the UNT Digital Library. This collection brings together articles, papers, presentations, books, chapters, reviews, academic posters, artwork, and other scholarly and creative works and makes them readily accessible to showcase UNT's research and creative achievements to a worldwide audience. The UNT Scholarly Works collection also serves as the open access repository for UNT."

Based on the context above, complete the following prompt by answering 'In scope:' for the given above 'Document:' Answer 'True' if it is in scope of UNT repository else answer 'False' Here are few examples:

Document : ABSTRACT Cost/Benefit Analysis and Ad Valorem Tax Benefits of Oil and Gas Drilling in the DFW Barnett Shale of Urban and Suburban North Texas John S. Baen Ph.D., University of North Texas. Key words/concepts: oil and gas valuation methods, need for education by real estate related professionals, planners and city administrators, city drilling ordinances, land-use efficiencies. This is a draft.

In scope : True

Document : Embedded Systems Design CSCE 3612, Section 001 and 002 Spring 2016 Class Timings: Tuesday and Thursday, 11:30 AM 12:50 PM, NTDP B142 Instructor: Robin Pottathuparambil Office Hours: Tuesday and Thursday 5:00 PM 6:00 PM or by appointment Instructional Assistant: Evan Rodrigues

In scope: False

Document : **(Document)**

In scope :

Chain of Thoughts

e. Chain-of-thoughts prompting:

Complete the following prompt by answering 'In scope:' for the given above 'Document:' Answer 'True' if it is in scope of UNT repository else answer 'False' Here are few examples:

Document : ABSTRACT Cost/Benefit Analysis and Ad Valorem Tax Benefits of Oil and Gas Drilling in the DFW Barnett Shale of Urban and Suburban North Texas John S. Baen Ph.D., University of North Texas. Key words/concepts: oil and gas valuation methods, need for education by real estate related professionals, planners and city administrators, city drilling ordinances, land-use efficiencies. This is a draft.

In scope : This is a manuscript which is among either articles, papers, presentations, books, chapters, reviews, academic posters, artwork, or other scholarly and creative works submitted to UNT. So this document is True

Document : Embedded Systems Design CSCE 3612, Section 001 and 002 Spring 2016 Class Timings: Tuesday and Thursday, 11:30 AM 12:50 PM, NTDP B142 Instructor: Robin Pottathuparambil Office Hours: Tuesday and Thursday 5:00 PM 6:00 PM or by appointment Instructional Assistant: Evan Rodrigues

In scope : This is just a course outline. This document does not belong to articles, papers, presentations, books, chapters, reviews, academic posters, artwork, or other scholarly and creative works submitted to UNT. Hence this document is False

Document : **(Document)**

In scope :

Figure 2: d.One-shot prompting with context e. Chain of thoughts prompting

surpassed the performance of some supervised, fine-tuned models on specific tasks. These breakthroughs catalyzed the development of numerous subsequent large language models such as Mistral [15], phi3 [1], Llama [35], PaLM [9], etc which demonstrated unique capabilities. Subsequent research introduced advanced prompting techniques such as Chain-of-Thought (CoT) [38], Tree-of-Thoughts (ToT) [39], and Self-consistency [37], which significantly enhanced the reasoning capabilities of LLMs.

In this work, We have used Llama2, Phi3, Mistral and Gemini as Language models and prompting techniques such as Zero-shot prompting, One-shot prompting and Chain-of-thought prompting.

3 METHODS

3.1 LLMs

We experiment with 4 LLM models namely Llama2, Phi3, Mistral, and Gemini.

3.1.1 Llama2. Llama 2 [35] is a family of transformer-based autoregressive large language models developed by Meta AI in 2023, which serves as the successor to the original Llama model. Detailed in the Llama 2 [35] research paper, these models were trained on 2 trillion tokens and are available in a scale ranging from 7 billion to 70 billion parameters. Notably, Llama 2 offers a context length

of 4096 tokens, which is double that of its predecessor, Llama 1. Additionally, the Llama 2 chat models were fine-tuned using reinforcement learning from human feedback, enhancing their conversational capabilities. For Llama model, we have used the chat models of Llama2-7B[5] model. We use the chat models since this model have been finetuned using Reinforcement learning with human feedback (RLHF).

3.1.2 Mistral. Mistral [15] is a prominent large language model introduced by Mistral AI in October 2023. It is a decoder-based model, mirroring the structure of the decoder block in the transformer architecture. According to the original Mistral research, the model supports a context length of 8192 tokens and is available with 7 billion parameters. Additionally, the Mistral attention block features 32 distinct heads, enhancing its capability to process and generate text. This design emphasizes efficiency and scalability, making Mistral a significant advancement in the field of natural language processing. For Mistral, we have used the *Mistral-7B-Instruct-v0.1*[6]. we also set the temperature value to 0.0001.

3.1.3 Phi3. Phi 3 [1] is a transformer-based model developed by the Microsoft Research team with a next-word prediction objective. These models are available in 2 different modes which are phi-3(7B parameters) small model and phi-3 medium model (14 B parameters). According to their research, Phi 3 small was trained on 3.3 trillion tokens and Phi 3 medium was trained on 4.8 trillion tokens. Notably, this model has not been aligned through reinforcement learning from human feedback nor has it undergone instruction fine-tuning, setting it apart from other models that often incorporate these additional training steps. We use the Phi3 mini version model with 128k context length[7].

3.1.4 Gemini. Gemini [34] is a multimodal language model developed by Google AI in 2024. This model is available in three distinct versions: Ultra, Pro, and Nano. Gemini supports a context window of up to 128,000 tokens, making it highly versatile for extended inputs. Built on a combination of Transformers and Mixture of Experts (MoE) architecture, Gemini exemplifies advanced design principles in AI. As a multimodal language model, it can process a variety of inputs, including text, video, and code, showcasing its broad applicability across different domains and tasks.

3.2 Prompting methods

In this paper, we use three types of prompts namely, Zero-shot prompting, one-shot prompting and Chain-of-thoughts prompting. Figures 1 and 2 contains examples for each prompts for UNT.edu dataset.

3.2.1 Zero-shot prompting. [5]:In this method, LLMs are tasked with completing a given objective solely based on a textual prompt, without the benefit of illustrative examples or demonstrations.

Zero-shot prompting without task description: In this method, the language model is instructed to perform a task without any description of the dataset. For this task, we need to instruct the LLM to classify if the particular document comes in the scope of a repository or not.

For this type of prompt, we instruct the LLMs to identify if the particular document is in the scope of UNT.edu, michigan.gov, or texas.gov and answer 'True' if it is in scope and 'False' if it is out-of-scope.

Zero-shot prompting with task description: In this method, we instruct the LLM to classify if the particular document comes in the scope of the UNT scholarly repository or not along with some description about the repository.

We include a context about the dataset along with the instructions. For UNT.edu we include the definition *"UNT Scholarly Works is an institutional repository of items contributed by the UNT Community and hosted in the UNT Digital Library. This collection brings together articles, papers, presentations, books, chapters, reviews, academic posters, artwork, and other scholarly and creative works and makes them readily accessible to showcase UNT's research and creative achievements to a worldwide audience. The UNT Scholarly Works collection also serves as the open access repository for UNT".*

For texas.gov the context about the Texas state publications is given as *"The Texas State Publications collection is composed on documents published by state agencies in Texas. This collection includes annual reports, research reports, white papers, newsletters, brochures, and budgets. It does not include blank forms, announcements, applications, receipts and most financial records, personal records, or FOIA requests. The audience for the Texas State Publications collection is the public with the goal of providing these documents to the public for research, education, or general scholarship."*

The context about michigan.gov is given as *"This collection contains a variety of public and/or published information from the executive, judicial and legislative branches of Michigan state government, excluding those materials defined in the other collections on this site. In general, this collection contains documents that convey information on agricultural, educational, historical, social, economic, political, environmental, judicial, cultural and health related topics specific to Michigan. "*

3.2.2 One-shot prompting. [5]: In this approach, the prompt includes a representative example for each relevant class, providing contextual information to guide the model's response.

One-shot prompting without task description: In this method, the LLM is only given examples with labels along with the document for which the scope (label) needs to be identified.

In one-shot prompting, we include one example per class and label each example. For UNT.edu we include a positive example *"ABSTRACT Cost/Benefit Analysis and Ad Valorem Tax Benefits of Oil and Gas Drilling in the DFW Barnett Shale of Urban and Suburban North Texas John S. Baen Ph.D., University of North Texas. Key words/concepts: oil and gas valuation methods, need for education by real estate related professionals, planners and city administrators, city drilling ordinances, land-use efficiencies. This is a draft."* and a negative example *"Embedded Systems Design CSCE 3612, Section 001 and 002 Spring 2016 Class Timings: Tues- day and Thursday, 11:30 AM 12:50 PM, NTDP B142 Instructor: Robin Pottathuparambil Office Hours: Tuesday and Thursday 5:00 PM 6:00 PM or by appointment Instructional Assistant: Evan Rodrigues"* and provide the labels as 'True' or 'False' below each example.

[5]Llama-7B - https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
[6]Mistral-7B - https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1
[7]Phi3-https://huggingface.co/microsoft/Phi-3-mini-128k-instruct

For michigan.gov we include the positive example *"PURCHASE OF FISH BY STATE AGENCIES AND LOCAL UNITS OF GOVERNMENT Act 183 of 1987 AN ACT to regulate the purchase of or contracting for the purchase of certain fish products by state agencies and certain local units of government ."* and a negative example *"SAMPLE - WIDA Paper Screener Order Form This order form is for the WIDA Paper Screener only . To submit order , complete order form and email to Jennifer Paul at paulj @ michigan.gov Your school will not receive an invoice for this order. The invoice will be sent to the Michigan Department of Education . Ship to Organization : School of Michigan Street Address : 1000 Michigan St. City : City of Michigan State : Michigan ZIP : 48909 Contact Name : Iwanna Screener E-mail : iwannascreener @ michigan.gov Phone : 517-000-000 KITS QUANTITY PRICE PER UNIT SUBTOTAL."*.

For texas.gov we include the positive example *"The Value of Distributed Photovoltaics to Austin Energy and the City of Austin This report was prepared as part of a response to SOLICITATION NUMBER: SL04300013 Study to Determine Value of Solar Electric Generation To Austin Energy Prepared for: Austin Energy Prepared by: Clean Power Research, L.L.C. Thomas E. Hoff Richard Perez Gerry Braun Michael Kuhn Benjamin Norris Final Report March 17, 2006 Executive Summary"* and a negative example *"LEGISLATIVE BUDGET BOARD Austin, Texas FISCAL NOTE, 80TH LEGISLATIVE REGULAR SESSION May 4, 2007 TO: Honorable John Whitmire, Chair, Senate Committee on Criminal Justice FROM: John S. O'Brien, Director, Legislative Budget Board IN RE: HB44 by Hodge (Relating to the restoration of good conduct time forfeited during a term of imprisonment.), As Engrossed No significant fiscal implication to the State is anticipated."*.

One-shot prompting with task description: In this method, the LLM is given examples along with a description of the task used in Zero-shot prompting to give more context to the model for predicting the scope of the given document. Here we use the same prompt that was used for One-shot without task description but include the description used in Zero-shot with context above the example set. Refer figure 2 for exact prompt format.

3.2.3 Chain-of-thought. [38]: Chain-of-thought (CoT) is a prompting technique that enhances LLMs' complex reasoning abilities by augmenting few-shot prompts with explicit explanations for each example document's classification. In this work, we incorporate document type as a rationale for classification, enabling the LLM to explicitly elicit the scope of the target document.

We used the same examples used in One-shot prompting. We ignore the context in this prompt as the rationale used under the documents includes the context of the dataset. However, we included an explanation under the examples along with labels. Since the LLMs are good at identifying the document type we give the document type as a reason for a document to be in-scope or out-of-scope of a repository.

For UNT.edu we give explanations as *"This is a manuscript which is among either articles, papers, presentations, books, chapters, reviews, academic posters, artwork, or other scholarly and creative works submitted to UNT. So this document is True"* and *"This is just a course outline. This document does not belong to articles, papers, presentations, books, chapters, reviews, academic posters, artwork, or other scholarly and creative works submitted to UNT. Hence this document is False"*.

For michigan.edu we include explanations such as *"This document is an order form which is among Forms, Announcements, Applications, Receipts and most financial records, Personnel records, Travel vouchers, FOIA request documentation RFP documentation, project documentation, Webpages with title or subject designations that contain only generalized text and links to related resources but no identifiable 'documents'. Hence answer is 'False'"* and *"This is a statute draft state government document enacted by michigan government which is NOT among Forms, Announcements, Applications, Receipts and most financial records, Personnel records, Travel vouchers, FOIA request documentation RFP documentation, project documentation, Webpages with title or subject designations that contain only generalized text and links to related resources but no identifiable 'documents'. Hence answer is 'True'"*

For texas.gov we include explanations such as *"This document is a report on solar power produced by the State of Texas, which is among collection that includes annual reports, research reports, white papers, newsletters, brochures. Hence answer is 'True'"* and *"This document is about a budget session (not a report but financial record) which is NOT among annual reports, research reports, white papers, newsletters, brochures. Hence answer is 'False'"*

All the explanations given above are to the documents used in the One-shot prompting discussed above.

4 DATASET AND EXPERIMENTS

We experiment with 4 types of prompts on all 4 LLMs. The experiments were conducted on 3 different datasets.

4.1 Dataset

In this paper, we have used three scholarly datasets namely unt.edu, texas.gov and michigan.gov. We check the performance of models on these datasets to find the best model for data-scarce setting.

UNT.edu: The first dataset was built from a web archive containing scholarly works from the University of North Texas (unt.edu). This archive was captured in May 2017, during a bi-yearly crawl of the university's website by the UNT Libraries. The archive contains 92,327 PDFs, representing only 3% of the total 3,141,886 URIs (web addresses). To assess the archive's relevance for a specific repository, researchers sampled 2,000 documents. They found that 445 (22%) were relevant, while the remaining 1,555 (78%) were not.

Michigan.gov: This dataset was created from web archives collected by the Library of Michigan using the Archive-It service available from the Internet Archive. This collection of websites has been collected since 2006 [8] The dataset was crawled between 2010 and 2023 and was generated from PDF files extracted from WARC files that were downloaded from the Archive-It collection. A total of 31,102,008 URLs are present in the entire web archive with PDF content (identified as URIs with an MIMETYPE of application/pdf) representing 808,400 of those URLS. After removing duplicate, malformed, or blank PDFs, a total of 77,747 documents were left. A random selection of 2,000 documents were sampled for labeling, 1834 documents (92%) were identified as being of interest for the Michigan State documents collection and 166 (8%) not being of interest.

[8] Michigan Government Web Collection - https://www.archive-it.org/collections/418

		UNT.edu		Michigan.gov		Texas.gov	
Llama2-7B	**ZS w/o TD**	0.5238	0.3929	0.5221	0.3967	0.4846	0.3625
	ZS with TD	0.5595	0.4900	0.5050	0.3560	0.5352	0.4956
	OS w/o TD	0.4277	0.3865	0.5437	0.4465	0.5010	0.5004
	OS with TD	0.6488	0.6394	0.5485	0.5397	0.5465	0.5435
	COT	**0.769**1	**0.748**0	**0.6125**	**0.6034**	**0.6762**	**0.6588**
Mistral-7B	**ZS w/o TD**	0.6488	0.6328	0.5050	0.37615	0.5469	0.5391
	ZS with TD	0.6845	0.6773	**0.5252**	**0.4481**	0.5641	0.5531
	OS w/o TD	0.5398	0.4032	0.4314	0.3466	0.5041	0.4563
	OS with TD	0.5535	0.2574	0.5050	0.3444	0.5762	0.5706
	COT	**0.7763**	**0.7738**	0.5101	0.3994	**0.6364**	**0.6278**
Phi3	**ZS w/o TD**	0.6011	0.6000	0.5027	0.4390	0.4966	0.4834
	ZS with TD	0.6795	0.6706	**0.62**61	**0.6153**	0.5566	0.5221
	OS w/o TD	0.4969	0.4629	0.5756	0.5411	0.5042	0.5031
	OS with TD	0.6130	0.6129	0.5477	0.5066	0.4962	0.4731
	COT	**0.7857**	**0.7846**	0.5625	0.5607	**0.606**	**0.604**
Gemini	**ZS w/o TD**	0.8630	0.8628	0.5757	0.5024	0.4866	0.4451
	ZS with TD	0.7857	0.7854	0.6717	0.6521	0.6448	0.6404
	OS w/o TD	0.5059	0.3563	0.6363	0.6278	0.5476	0.5431
	OS with TD	0.7619	0.7529	**0.6919**	**0.6819**	0.6657	0.6556
	COT	**0.8928**	**0.8809**	0.6464	0.6420	**0.7364**	**0.7321**
BERT	**supervised**	0.8511	0.8511	0.7121	0.7094	0.7720	0.7709
RoBERTa	**supervised**	**0.8690**	**0.8609**	**0.7423**	**0.7401**	**0.8204**	**0.8200**

Table 1: Performance of LLMs with Zero-shot without task description (ZS w/o TD), Zero-shot with task description (ZS with TD), One-shot without task description (OS w/o TD), One-shot without task description (OS with TD), Chain-of-Thought (CoT) prompts for all datasets. These results are compared with BERT and RoBERTa fine-tuned with 50 examples per class.

Texas.gov: This dataset was built from a web archive of the Texas government websites crawled between 2002 and 2011 which is stored by the UNT digital library. It contains 1,752,366 PDFs, representing 6.7% of the total 26,305,347 web addresses (URIs). Out of a 2,000-document sample, only 136 (7%) were relevant for a particular repository.

4.2 Test data

Since the scope of this paper is to use only prompts and not to finetune the model, we do not require any training data. However, for the sake of One-shot prompting and Chain-of-thought prompts which require one example from each class, we randomly select one example from the training data for each dataset. We have 168 samples from the UNT.edu dataset, 198 samples from the Michigan.gov dataset, and 250 samples from the texas.gov dataset where all these samples are balanced between both classes.

4.3 Experimental setup

We evaluate the performance of models when used with different prompts mentioned in section 3.2. We use the first 100 tokens to evaluate the models. We conduct a comparative analysis of our proposed methods against supervised learning baselines employing fine-tuned BERT and RoBERTa models. Moreover, To assess the impact of document length on model performance, we evaluated the models using two approaches: the first X words and the first X plus last X words, employing Chain-of-thought prompting, where X $\in$ {100, 300, 500, 700, 900}. The main reason for choosing X in those subsets is that we know most of the keywords that are useful for classification fall under these ranges [28] and the reason for choosing a Chain-of-thoughts is that it gives us higher performance than other prompts. The results presented in Figures 3, 4, and 5 indicate that performance varies depending on the specific model and dataset utilized.

Furthermore, to assess model sensitivity to prompt variations, evaluations were conducted using prompts containing three distinct sets of examples. Given the established effectiveness of "Chain-of-thought" and "One-shot prompting with context," these two prompting methods were chosen for the evaluation. Once again, only the initial 100 tokens of each document were utilized.

To investigate the influence of increasing exemplar quantity on Chain-of-Thought prompting, we compare One-shot, Two-shot, and Five-shot CoT methods. The performance of these approaches is evaluated against supervised learning baselines. Given its superior performance on other tasks, we employ the Gemini model for our experiments.

5 RESULTS AND ANALYSIS

In this section, we first compare the performance of all models on different types of prompts. Then, we analyze the performance of Chain-of-Thoughts prompts on all models with different document lengths. Finally, we analyze how Chain-of-Thoughts prompting performs when the example documents in the prompts are changed.

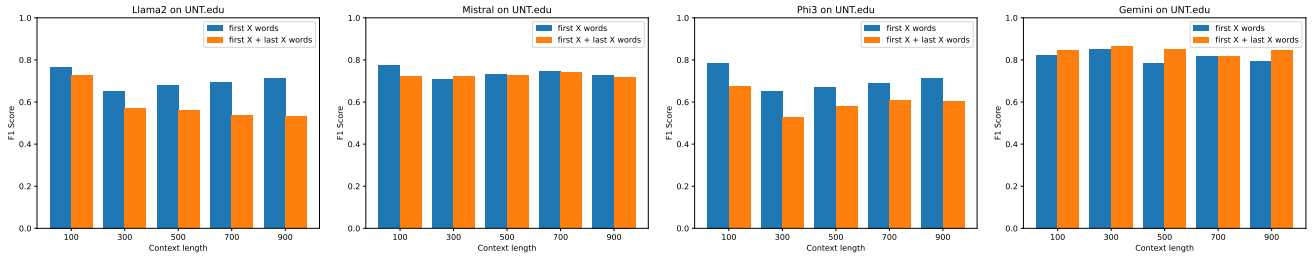

Figure 3: Performance of models using first X words and first and last X words on UNT.edu dataset on Llama, Mistral, Phi3, Gemini (Left to right)

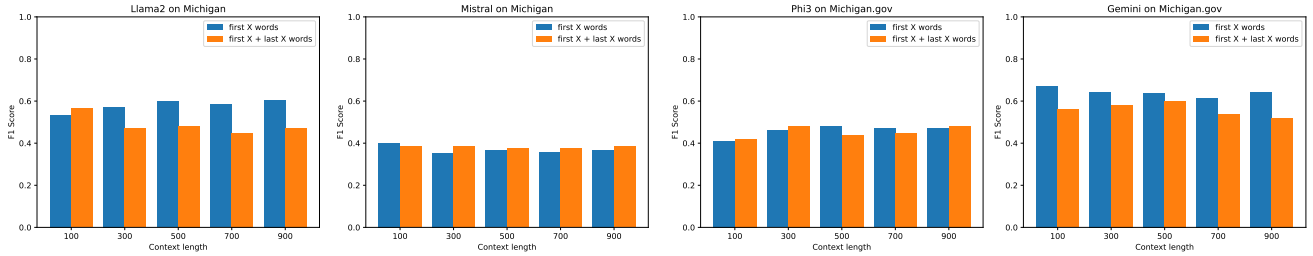

Figure 4: Performance of models using first X words and first and last X words on Michigan.gov dataset on Llama, Mistral, Phi3, Gemini (Left to right)

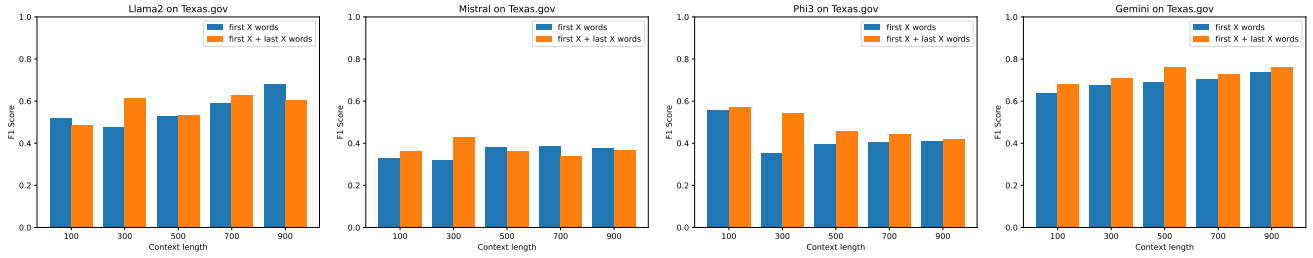

Figure 5: Performance of models using first X words and first and last X words on texas.gov dataset on Llama, Mistral, Phi3, Gemini (Left to right)

Based on the results presented in Table 1 across three distinct datasets, we conclude that employing Chain-of-thought prompts significantly enhances the performance of large language models (LLMs) in low-resource settings. Furthermore, our experiments indicate that the Gemini model consistently outperforms all other models tested in the context of document classification. For example, in all three datasets, it is evident that the Chain-of-thought prompts yield the highest accuracy and F1-scores (except for Phi3 and Mistral model on michigan dataset which could be due to the nature of dataset) compared to other prompting methods for every model. When comparing results model-wise, the Gemini model yields 89%, 69%, and 73% accuracy on the UNT.edu, Michigan.gov, and Texas.gov datasets, respectively, which is significantly higher than any other model. These results allow us to develop further conclusions, which are stated below:

5.1 Analysis of prompts

The performance of the models on different types of prompts can be seen in Table 1. Based on the table we provide the following analysis.

5.1.1 Adding task description improves performance. : Our primary objective was to leverage the ability of large language models (LLMs) to classify documents in low-resource settings. When using Zero-shot prompting, the models were able to classify some documents, although with low accuracy across all datasets, except the Gemini model on the UNT.edu dataset. This exception might be attributed to the specific characteristics of the dataset or the training data of the Gemini model. Generally low accuracy in Zero-shot prompting can be attributed to the difficulty LLMs face in converging to a single domain without explicit instructions, due to their training across multiple domains. However, when additional context, such as task descriptions, was provided for Zero-shot learning, a slight

			Example-1		Example - 2		Example - 3		Sensitivity of F1	
			Acc	F1	Acc	F1	Acc	F1	Mean	s.d.
UNT.edu	**Gemini**	**One shot**	0.8273	0.8234	0.5476	0.5174	0.6071	0.5655	0.6593	0.0090
		COT	0.8928	0.8923	0.8809	0.8795	0.7797	0.7796	0.6874	0.0522
	Llama - 7B	**One shot**	0.6488	0.6394	0.5670	0.5246	0.5644	0.5334	0.4533	0.0757
		COT	0.7690	0.7480	0.7086	0.7016	0.6582	0.6501	0.5363	0.1090
Michigan.gov	**Gemini**	**One shot**	0.6919	0.6819	0.6060	0.6027	0.6978	0.6777	0.6541	0.0446
		COT	0.6464	0.6420	0.5757	0.5468	0.7070	0.7065	0.6318	0.0803
	Llama - 7B	**One shot**	0.5400	0.5300	0.4947	0.4681	0.4734	0.3997	0.4659	0.0652
		COT	0.6125	0.6034	0.6721	0.6709	0.5839	0.5717	0.6153	0.0507
Texas.gov	**Gemini**	**One shot**	0.6600	0.6500	0.6600	0.6600	0.6700	0.6680	0.6354	0.1646
		COT	0.7360	0.7321	0.7000	0.7000	0.6300	0.6300	0.8505	0.0617
	Llama - 7B	**One shot**	0.5400	0.5400	0.4800	0.4000	0.5400	0.4200	0.5658	0.0639
		COT	0.6762	0.6588	0.5300	0.5000	0.5000	0.4500	0.6999	0.0490

Table 2: Performance of Llama2 and Gemini models on each dataset by changing the examples in One-shot with context and Chain-of-thoughts prompts. The relative standard deviation of accuracy and F1-score for each set of examples to a model for a dataset is given last column. NOTE: The Example-1 column takes values from table 1

	UNT.edu		Michigan.gov		Texas.gov	
One-shot CoT	0.8928	0.8809	0.6464	0.6420	0.736	0.7321
Two-shot CoT	0.8383	0.8383	0.7121	0.7120	0.652	0.6495
Five-shot CoT	**0.9285**	**0.9284**	**0.7323**	**0.7264**	**0.7840**	**0.7816**
BERT	0.8511	0.8511	0.7121	0.7094	0.7720	0.7709
RoBERTa	**0.8690**	**0.8609**	**0.7423**	**0.7401**	**0.8204**	**0.8200**

Table 3: Comparative analysis of LLM performance when prompted with K examples versus supervised learning approaches.

increase in classification performance was observed. The most significant improvements in F1- scores were 25.6% on UNT.edu with Llama 2, 41.1% on Michigan.gov with Phi 3, and 45.4% on Texas.gov with Gemini.

When performing One-shot learning to evaluate the classification ability of LLMs, we observed an overall slight decrease in performance. This decrease could be attributed to the lack of context for the repositories, as a divergent example in the repository might mislead the models during convergence. To address these challenges, it is beneficial to include additional context for One-shot prompting. Upon doing so, we observed a slight increase in performance for most models across all three datasets.

The main disadvantage of One-shot learning with context prompts is that we only provide binary labels (true/false or in-scope/out-of-scope) for the examples without indicating the rationale behind these labels. This lack of explanation can confuse the models, similar to how humans sometimes struggle with document classification based solely on a collection description. However, when we provide additional explanation for the answers (i.e., Chain-of-thought prompting), we observe a significant increase in model performance across all three datasets, except the Phi 3 and Mistral models on the Michigan dataset. This performance improvement is likely due to the further explanation, which helps the models converge towards a specific domain.

Overall we observe that providing more context about the task description helps the models increase their classification ability.

5.1.2 Document type identification. : Among all the prompts, Chain-of-Thoughts prompting has been shown to perform well in most cases. Apart from adding context, document type determines if the current document is in-scope or out-of-scope of a collection. So, including the document type as a reason in prompts improves the performance. The positive impact of structural features on document type classification, as highlighted by [28], is evident in our results.

5.2 Analysis on Length of the documents

To understand the effect of the length of a document on the performance of the model, we evaluated models based on first X words and first X and last X words with Chain-of-thoughts prompting. The results in figures 3, 4, 5 show that the performance depends on the model and dataset used.

In the UNT.edu dataset, the models Llama, Mistral, and Phi3 demonstrate superior performance when utilizing the first X words compared to the combination of the first X and last X words. This may suggest that these models benefit from using the initial keywords or context for enhanced classification. Conversely, the Gemini model performs better with the combination of the first and last X words, indicating that it may leverage a broader context or identify more relevant keywords within the last X words for classification. For the Michigan.gov dataset, the models Llama, Gemini, and Phi3 also achieve higher accuracy with the first X words than with the combination of the first X and last X words. This may imply that these models effectively utilize the initial keywords for

improved classification. In contrast, Mistral shows slightly better performance with the combination of the first and last X words, suggesting a potential convergence advantage in its classification approach. In the Texas.edu dataset, Llama and Mistral again excel with the first X words compared to the combination of the first and last X words, indicating that these models may benefit from the initial context for classification. However, the Gemini and Phi3 models perform better with the combination of the first and last X words, likely due to their ability to leverage a more comprehensive context or to identify keywords in the last X words. Overall, we found that model performance is not solely dependent on document length; rather, it varies across different contexts. For instance, the Gemini model performs better with the combination of the first X and last X words on the UNT.edu and Texas.gov datasets, but this is not the case for Michigan.gov. This suggests that model performance is influenced not only by document length but also by the nature of the repository and the effectiveness of the context provided, such as the presence of relevant keywords in either the initial or final segments of the text.

5.3 Effect of change of examples in One-shot and Chain-of-Thoughts

To understand the effect of change of examples in One-shot with context and Chain-of-thoughts, we try evaluating by replacing current examples with two more examples, for both positive and negative classes. To quantify this effect we compute the relative standard deviation in terms of percentage for F1-Score with all 3 examples.

From table 2 based on the standard deviation values, we observed that UNT.edu dataset is very sensitive to the change of examples in comparison to other datasets.

5.4 Effect of increasing the number of examples in CoT prompting

As demonstrated in Section 5.3, careful selection of exemplar documents is critical for effective prompting. To mitigate the challenge of manual exemplar selection, we propose incorporating a set of K examples for each class within the prompt. This approach increases the possibility of including optimal exemplars and provides the model with additional contextual information about the task. As anticipated, expanding the number of exemplars enhanced the model's classification capabilities, as evidenced in Table 3. Notably, the performance of our proposed method approaches that of fine-tuned models.

6 CONCLUSION

Our work demonstrates the effectiveness of prompting techniques for classifying documents within a collection's scope, without model fine-tuning. We found that "Chain-of-thoughts" prompting, incorporating document type as the rationale for belonging to a class, consistently outperformed other methods. Additionally, including context about in-scope documents within the prompts further improved model performance. These findings suggest that carefully crafted prompts can significantly enhance the ability of Large Language Models (LLMs) to identify relevant documents. However, our study also highlights the sensitivity of prompt performance to specific example documents within One-shot and Chain-of-thoughts prompts. This sensitivity suggests the importance of identifying optimal example sets for each model and dataset, which presents a promising avenue for future research. Future research could delve deeper into prompting techniques that utilize library metadata and expand experiments to web archives in multiple languages. Additionally, exploring the application of these techniques to fine-tuning and assessing their practical usability would be valuable.

ACKNOWLEDGEMENTS

We thank the Institute of Museum and Library Services for support from a grant which supported the research in this study. We also thank our anonymous reviewers for their insightful feedback and comments.

REFERENCES

[1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Yen-Chun Chen, Yi-Ling Chen, Parul Chopra, Xiyang Dai, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Mei Gao, Min Gao, Jianfeng Gao, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Ce Liu, Mengchen Liu, Weishung Liu, Eric Lin, Zeqi Lin, Chong Luo, Piyush Madan, Matt Mazzola, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Xin Wang, Lijuan Wang, Chunyu Wang, Yu Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Haiping Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Jianwei Yang, Ziyi Yang, Yifan Yang, Donghan Yu, Lu Yuan, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. arXiv:2404.14219 [cs.CL] https://arxiv.org/abs/2404.14219

[2] Brenda Reyes Ayala and Cornelia Caragea. 2014. Towards building a collection of web archiving research articles. *Proceedings of the American Society for Information Science and Technology* 51, 1 (2014), 1–5.

[3] Jefferson Bailey. 2017. Twitter Post. https://twitter.com/jefferson_bail/status/867808876917178368.

[4] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. *Advances in neural information processing systems* 13 (2000).

[5] 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 (2020), 1877–1901.

[6] Cornelia Caragea, Jian Wu, Sujatha Gollapalli, and C Giles. 2016. Document type classification in online digital libraries. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30. 3997–4002.

[7] Maria Fernanda Caropreso, Stan Matwin, and Fabrizio Sebastiani. 2001. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. *Text databases and document management: Theory and practice* 5478, 4 (2001), 78–102.

[8] Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2023. Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307* (2023).

[9] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research* 24, 240 (2023), 1–113.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[11] C. Dooley and G. Thomas. 2019. The library of congress web archives: Dipping a toe in a lake of data. *https: //blogs.loc.gov/thesignal/2019/01/ the-library-of-congress-web-archivesdipping-a-toe-in-a-lake-of-data/* (2019).

[12] Nathaniel T Fox, Mark E. Phillips, and Hannah Tarver. 2020. Programmatic Extraction of 'Documents' from Web Archives: Identifying Document Characteristics from Content Selector Interviews. https://digital.library.unt.edu/ark:/67531/metadc1757659/.

[13] Yoav Goldberg. 2016. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research* 57 (2016), 345–420.

[14] Cathy Nelson Hartman and Coby Condrey. 2004. TRAIL: From Government Information Locator Service to Electronic Depository Program for Texas State Publications. *DttP: Documents to the People* 32 (2004), 22–27. Issue 2.

[15] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. arXiv:2310.06825 [cs.CL] https://arxiv.org/abs/2310.06825

[16] Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*. Springer, 137–142.

[17] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188* (2014).

[18] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).

[19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[20] Julien Masanès. 2005. Web archiving methods and approaches: A comparative study. *Library trends* 54, 1 (2005), 72–90.

[21] Andrew McCallum, Kamal Nigam, et al. 1998. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, Vol. 752. Madison, WI, 41–48.

[22] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196* (2024).

[23] K. R. Murray and I. K. Hsieh. 2008. Archiving Web-published materials: A needs assessment of librarians, researchers, and content providers. *Government Information Quarterly* 25 (2008), 66–89. Issue 1.

[24] Alexandros Ntoulas, Junghoo Cho, and Christopher Olston. 2004. What's new on the Web? The evolution of the Web from a search engine perspective. In *Proceedings of the 13th international conference on World Wide Web*. 1–12.

[25] OpenAI. 2022. Introducing ChatGPT. https://openai.com/blog/chatgpt

[26] 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 (2022), 27730–27744.

[27] Tevor Owens. 2019. The Library of Congress Web Archives: Dipping a Toe in a Lake of Data. https://blogs.loc.gov/thesignal/2019/01/the-library-of-congress-web-archives-dipping-a-toe-in-a-lake-of-data/.

[28] Krutarth Patel, Cornelia Caragea, and Mark Phillips. 2020. Dynamic classification in web archiving collections. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 1459–1468.

[29] Krutarth Patel, Cornelia Caragea, Mark E Phillips, and Nathaniel T Fox. 2020. Identifying documents in-scope of a collection from web archives. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*. 167–176.

[30] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

[31] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[32] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* 21, 140 (2020), 1–67.

[33] Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)* 34, 1 (2002), 1–47.

[34] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[35] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs.CL] https://arxiv.org/abs/2307.09288

[36] 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 (2017).

[37] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171* (2022).

[38] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[39] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems* 36 (2024).

[40] Ye Zhang and Byron Wallace. 2015. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820* (2015).

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009