

SANTA CLARA UNIVERSITY

Department of Computer Science and Engineering

Date: August 6, 2024

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY
SUPERVISION BY

Jingsen Wang

ENTITLED

A Step Towards Automated Ethical Analysis in Journalism: Measuring LLMs' Performance in Extracting Sourcing Information

BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE OF


Master of Science in Computer Science and Engineering


Yi Fang (Aug 14, 2024 11:48 PDT)

Thesis Advisor: Dr. Yi Fang


Yuhong Liu (Aug 14, 2024 13:56 PDT)

Thesis Reader: Dr. Yuhong Liu



Chair of Department: Dr. Silvia Figueira

SANTA CLARA UNIVERSITY

MASTER'S THESIS

**A Step Towards Automated Ethical
Analysis in Journalism: Measuring
LLMs' Performance in Extracting
Sourcing Information**

Author:

Jingsen WANG

wjingsen@gmail.com

Supervisor:

Dr. Yi FANG

yfang@scu.edu

Computer Science and Engineering

August 6, 2024

“The first principle is that you must not fool yourself and you are the easiest person to fool.”

Richard P. Feynman

Acknowledgements

I would like to extend my heartfelt thanks to the five undergraduate students from the Markkula Center’s Spring 2024 internship class—Gigi Patmore, Kelly Perasso, Emily Hofstetter, Sarah El Shenawy, and Reagan Viens—for their diligent efforts in annotating news stories and developing the ground truth annotation checklist.

I am especially grateful to Subbu Vincent, director of Journalism and Media Ethics at the Markkula Center, for his invaluable guidance and support throughout this thesis.

I also want to express my sincere thanks to Dr. Yuhong Liu, my thesis reader, for her informative and detailed feedback that greatly improved this thesis.

Lastly, I am deeply thankful to my thesis advisor, Dr. Yi Fang, for his expertise in NLP, kindness, and insightful advice, which have greatly enriched this work.

Abstract

This thesis explores the potential of Large Language Models (LLMs) in automating the extraction of sourcing information from news articles, a crucial step towards enhancing transparency and ethical analysis in journalism. We evaluate the performance of two state-of-the-art LLMs, GPT-4 and Claude 3, in identifying and categorizing various source types across four diverse news articles. The thesis employs a zero-shot learning approach with two different prompt designs, assessing the models' ability to adapt to varying source structures and prompt instructions.

Our findings reveal that while LLMs show promise in extracting sourcing information, their performance varies significantly across different article types and source structures. The research highlights the complex interplay between prompt design, source types, and model performance, with both LLMs demonstrating strengths and limitations in handling diverse journalistic contexts. This thesis contributes to the growing body of work on AI in journalism by providing initial insights into the current capabilities of LLMs in sourcing analysis and outlining key areas for future research and development in automated ethical analysis of news content.

KEYWORDS: LLMs, NLP, Journalism, Journalism Ethics, Sources, Attributions

Table of Contents

Acknowledgements	ii
Abstract	iii
1 Introduction	1
1.1 Background	1
1.2 The Promise of Large Language Models	2
1.3 Vision for Transparent Sourcing	3
1.4 Challenges and Limitations	3
1.4.1 Limitations of LLMs:	3
1.4.2 Prompt Engineering Complexity:	4
1.4.3 Lack of Benchmarks:	4
1.4.4 Resource-Intensive Ground Truth Creation:	4
1.4.5 Scope Definition:	5
1.5 Thesis Objectives	5
1.6 Expected Outcome and Significance	6
2 Related Work	7
2.1 Large Language Models	7
2.1.1 In-Context Learning in LLMs	10
2.1.2 Hallucination in LLMs	12
2.2 Sourcing in Journalism	12
2.3 Automated Evaluation of News Articles	14
2.4 Summary	17

3	Our Approach	18
3.1	Selection of Large Language Models	18
3.2	News Article Dataset	19
3.3	Sourcing Information Extraction	19
3.4	Prompting Technique	20
3.5	Benchmark Creation: Human Annotation	21
3.6	Evaluation Methods	22
4	Experiments and Results Discussion	24
4.1	Experiment Setup	24
4.1.1	Data Preparation	24
4.1.2	LLM Analysis Process	24
4.1.3	LLM Prompt Strategies	25
	V0: Person-Only Prompts	25
	V1: Person and Organization Prompts	26
4.1.4	Evaluation Process	26
4.1.5	Experimental Parameters	29
	LLM Models:	29
	LLM Configuration:	29
	Similarity Scores Computation:	30
	Output Format:	31
4.1.6	Experimental Runs	31
4.2	Results Discussion	32
4.2.1	Consistency Across Runs	32
	Claude’s Performance	32
	GPT-4’s Performance	32
4.2.2	Performance Comparison	33
	LLM Recall	33
	Unique Discovery Rate	33

Name and Source Type Match Rates	34
Title and Association Match Rates	34
4.2.3 Article-Specific Observations	35
4.2.4 Comparison of Prompt Versions	37
Overall Performance Comparison	37
Key Observations	37
Analysis of Prompt Version Impact	39
5 Conclusion and Future Work	40
5.1 Limitations	40
5.2 Conclusion	41
5.3 Future Work	42
5.3.1 Enhanced Instruction Design	42
5.3.2 Larger and More Diverse Datasets	43
5.3.3 Real-Time Applications	43
5.3.4 On-demand Source Literacy tools for Everyday News Consumers	43
5.3.5 Addressing Ethical Considerations	44
5.3.6 Cross-Disciplinary Collaboration	44
5.4 Final Thoughts	44
References	46
A Appendix A	51
System Prompt - V0 - Persons Only	51
User Prompt - V0	52
System Prompt - V1 - Persons and Organizations	52
User Prompt - V1	53

List of Tables

4.1	Claude’s LLM Recall Consistency Across Three Runs	32
4.2	GPT-4’s LLM Recall Consistency Across Three Runs	32
4.3	Average LLM Recall Comparison Between Claude and GPT-4	33
4.4	Average Unique Discovery Rate Comparison Between Claude and GPT-4	33
4.5	Average Name Match Rate and Source Type Match Rate . . .	34
4.6	Average Title Match Rate and Association Match Rate	35
4.7	Comparison of Average LLM Recall Between Prompt Versions 0 and 1	37

Dedicated to my dear family. To my dad, who supports me in pursuing my craziest dreams. To my mom, who believes in my strengths. To my little brother, who challenges me to go beyond my limitations. Without them, this work would not have been possible.

Chapter 1

Introduction

1.1 Background

In the digital age, where information spreads rapidly and news consumption patterns have shifted dramatically, the public's ability to assess the credibility of news sources has become increasingly critical (Coddington and Molyneux (2023)). This digital transformation of the media landscape, coupled with factors such as the spread of misinformation and political polarization, may have contributed to a decline in public trust towards media organizations (Gottfried and Liedke (2021)). Amidst these challenges, enhancing the transparency of journalistic practices, particularly in sourcing, emerges as a potential pathway to address the complex issue of media credibility and public trust.

Sourcing is a fundamental pillar of credible journalism (Kovach and Rosenstiel (2014)). It involves the process of obtaining information from various sources and transparently presenting them in news stories. Proper sourcing practices lend credibility to news articles, allowing readers to verify claims and understand the context of the information presented.

However, evaluating the quality and ethics of sourcing in news articles requires a certain level of domain knowledge in journalism (Coddington and Molyneux (2023)). This creates a significant barrier for the general public,

who may lack the necessary background to critically assess the sourcing practices employed in the news they consume. As a result, there is a growing need for tools and methodologies that can make sourcing information more transparent and accessible to the average reader.

1.2 The Promise of Large Language Models

In recent years, the development of Large Language Models (LLMs) has opened up new possibilities in various fields, including natural language processing and information extraction (Qin et al. (2023)). These models have demonstrated remarkable capabilities in understanding and generating human-like text, making them potentially powerful tools for analyzing and interpreting complex information.

Two key attributes of LLMs make them particularly promising for addressing the challenges of sourcing transparency in journalism:

1. **Efficient Information Extraction:** LLMs have shown exceptional ability in extracting relevant information from large volumes of text (Törnberg (2023), Goel et al. (2023)). This capability could be leveraged to automatically identify and categorize sourcing information within news articles.

2. **In-Context Learning:** One of the most exciting features of modern LLMs is their ability to perform in-context learning (Wang et al. (2024)). This allows these models to adapt to new domains and tasks without fine-tuning, essentially providing them with "new domain knowledge" on the fly (Min et al. (2022)). This flexibility could be crucial in adapting LLMs to the complex and varied sourcing practices used in journalism, including different types of source attribution, and different context of news stories.

1.3 Vision for Transparent Sourcing

Given the capabilities of LLMs, we envision a future where these technologies can be leveraged to make sourcing information in news articles more transparent and accessible to the general public. The core idea is to develop a system that can:

- Automatically extract and categorize sourcing information from news articles using LLMs.
- Present this information in a clear, easily understandable format for readers.
- Develop an Ethics Metrics based on the extracted sourcing information to evaluate the credibility and transparency of different news providers.

By making sourcing practices more visible and comprehensible to the average reader, we aim to empower the public to make more informed judgments about the credibility of news sources. This increased transparency could play a crucial role in rebuilding trust in media organizations that consistently demonstrate ethical sourcing practices.

1.4 Challenges and Limitations

While the potential of using LLMs for improving sourcing transparency is exciting, several significant challenges must be addressed:

1.4.1 Limitations of LLMs:

- **Lack of Interpretability:** Despite their impressive performance, the decision-making processes of LLMs often remain opaque, making it difficult to understand how they arrive at their conclusions (Zhang (2024), Singh et al. (2024)).

- **Hallucination:** LLMs can sometimes generate plausible-sounding but factually incorrect information, a phenomenon known as "hallucination"(Huang et al. (2023)). This is particularly concerning when dealing with sensitive journalistic content.
- **Reliability Issues:** As probabilistic models, LLMs may produce inconsistent results, raising questions about their reliability for critical tasks (Reiss (2023)).

1.4.2 Prompt Engineering Complexity:

Developing effective prompts for LLMs is a challenging and often iterative process(Gao et al. (2024)). The vast number of possible prompt combinations means that initial failures may not necessarily indicate a lack of capability in the LLM, but rather a need for further experimentation and refinement .

1.4.3 Lack of Benchmarks:

Currently, there are no established benchmarks for evaluating LLM performance in the specific task of identifying and assessing journalistic sourcing practices(Hendrycks et al. (2021)). This absence makes it difficult to compare different approaches and measure progress.

1.4.4 Resource-Intensive Ground Truth Creation:

Building a comprehensive, human-annotated dataset for training and evaluating LLMs on sourcing identification is a time-consuming and expensive process, given the complexity and nuance involved in journalistic sourcing practices.

1.4.5 Scope Definition:

Sourcing in journalism is a vast and complex subject (Schudson (2011)). Determining how to narrow this down into manageable, computable tasks for LLMs without losing essential nuances is a significant challenge.

1.5 Thesis Objectives

While the vision of leveraging LLMs to improve transparency of sourcing in public media is ambitious, this thesis aims to take the first step towards realizing this goal. The primary objective of this thesis is to conduct a foundational experiment assessing the capability of Large Language Models in extracting sourcing information from news articles, establishing a baseline for future research in this domain.

Specifically, this thesis seeks to:

- Develop and implement a minimal experimental framework to evaluate LLMs' ability to extract sourcing information from news articles.
- Test multiple prompt sets to ensure the results are not biased by a single prompt formulation.
- Evaluate at least two different LLMs (e.g., Claude 3 and GPT-4) to provide a comparative analysis of their capabilities.
- Select a manageable set of news articles that represents a range of topics, article types, and news sources, to serve as a diverse testing ground for the LLMs' sourcing extraction capabilities.
- Develop a human-annotated benchmark for comparison, acknowledging potential imperfections due to resource limitations.
- Design and implement quantitative metrics to objectively compare the performance of LLMs against the human-annotated benchmark.

1.6 Expected Outcome and Significance

By conducting these experiments, we expect to gain valuable insights into the potential of LLMs for extracting sourcing information from news articles. The results will provide a baseline understanding of LLMs' capabilities in this domain, highlight areas for improvement, and identify promising directions for future research.

The significance of this work lies in its potential to contribute to the broader goal of improving transparency in journalism. By exploring the use of cutting-edge AI technologies in this context, we are taking an important step towards developing innovative solutions that can empower readers and strengthen the foundations of informed democratic discourse.

Chapter 2

Related Work

Understanding the capabilities and limitations of LLMs in the context of NLP tasks is crucial for developing effective applications in various domains, including journalism. This chapter reviews existing literature on LLMs, their application in NLP tasks, the importance of sourcing in journalism, and previous efforts to automate the evaluation of news article sourcing.

2.1 Large Language Models

LLMs, such as GPT-3 and BERT have revolutionized the field of NLP by demonstrating remarkable performance across a wide range of tasks. These models leverage vast amounts of data and sophisticated architectures to understand and generate human-like text. Key studies in this area include:

- GPT-3: In this groundbreaking paper Brown et al. (2020) from OpenAI introduced GPT-3, a large-scale autoregressive language model with 175 billion parameters. The key contribution of this work was demonstrating that scaling up language models to such an unprecedented size enables them to perform few-shot learning - the ability to solve tasks given only a few examples or a natural language prompt, without any fine-tuning.

The authors extensively evaluated GPT-3 in zero-shot, one-shot, and few-shot settings across a wide range of NLP tasks and benchmarks. They found that GPT-3 achieved promising results in the zero- and one-shot settings, and in the few-shot setting its performance was sometimes competitive with state-of-the-art fine-tuned models, despite using far fewer task-specific training examples.

This work highlighted the potential of very large language models for task-agnostic, few-shot learning. It suggested that scaling up models and pretraining them on broad data might be a path towards more general and adaptable language systems that can perform a variety of tasks with minimal explicit supervision. GPT-3's few-shot capabilities sparked significant interest and follow-up work in the field on instruction tuning, in-context learning, and developing LLMs as general-purpose "foundations" for downstream tasks.

- GPT-4 OpenAI et al. (2024): In our exploration of automated sourcing analysis in journalism, we consider GPT-4, a state-of-the-art large language model developed by OpenAI. Released in March 2024, GPT-4 represents a significant leap in natural language processing capabilities. It demonstrates human-level performance on various professional and academic benchmarks, including scoring in the top 10% of test-takers on a simulated bar exam.

Particularly relevant to our thesis, GPT-4 exhibits enhanced reasoning abilities, improved factual accuracy, and more nuanced task comprehension compared to its predecessors. Its ability to handle longer text inputs and its reduced tendency for hallucination make it a promising tool for analyzing complex journalistic content. These advancements suggest potential benefits for tasks like identifying and categorizing source types in news articles.

However, like all AI models, GPT-4 has limitations, including potential biases and occasional errors in reasoning. These limitations underscore the importance of our comparative study, as we seek to understand the current capabilities and limitations of LLMs in the context of journalistic analysis and ethical considerations.

- BERT: Devlin et al. (2019) introduced BERT (Bidirectional Encoder Representations from Transformers), a novel approach for pre-training language representations.

The key innovation of BERT lies in its use of bidirectional training of a Transformer model, which allows the model to learn from both left and right context in all layers. This is in contrast to previous approaches like OpenAI GPT which were limited to a left-to-right architecture. BERT employs two pre-training tasks - Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) - that enable it to learn robust bidirectional representations.

The authors demonstrated that pre-trained BERT representations can be fine-tuned with just one additional output layer to achieve state-of-the-art performance on a wide range of NLP tasks, such as question answering, language inference, and text classification. BERT advanced the state of the art on 11 NLP tasks at the time.

BERT's bidirectional pre-training approach and its strong performance across a broad range of tasks made it immensely influential, marking a significant milestone in the use of pre-trained language models for NLP. It paved the way for further advancements in language model pre-training and fine-tuning, and established bidirectionality as a key ingredient in achieving robust language understanding in such models.

- Claude 3 Anthropic (2024): Introduced by Anthropic in March 2024, Claude 3 represents a significant advancement in LLM capabilities, particularly relevant to this thesis's focus on automated sourcing information extraction in journalism. As part of the Claude 3 family, Claude 3 Opus demonstrates state-of-the-art performance in complex language understanding tasks. Key features aligned with the thesis objectives include enhanced textual analysis capabilities, improved contextual understanding, and the ability to handle extensive context windows with robust recall. These attributes are valuable for identifying and categorizing various source types within news articles. Claude 3's advanced language processing capabilities position it as a promising tool for the zero-shot learning approach employed in this thesis. Anthropic's emphasis on responsible AI development in Claude 3, including efforts to mitigate biases, aligns with the ethical considerations inherent in analyzing journalistic content. As such, Claude 3 serves as a valuable benchmark for assessing the current state and future potential of LLMs in advancing automated ethical analysis in journalism.

2.1.1 In-Context Learning in LLMs

In-context learning (ICL) is a paradigm where LLMs learn to perform tasks based on a few examples provided in the input prompt, without requiring any gradient updates or fine-tuning. This capability is particularly valuable for extracting specific information, such as sourcing entities in news articles.

- Overview (Dong et al. (2023)): This survey provides a comprehensive overview of ICL in LLMs. The survey covers various techniques for improving ICL performance, including model warmup strategies and demonstration engineering strategies. It also discusses the factors influencing ICL performance and its potential applications.

We reference this work because our thesis leverages the ICL capability of LLMs to identify and extract essential entities related to sourcing in news articles. By providing detailed instructions and examples, we aim to guide LLMs in recognizing named and unnamed sources, their titles, and attributed statements. The survey by Dong et al. offers valuable insights into the current state of ICL research and its potential applications, which aligns with our goal of using LLMs for automated ethical analysis in journalism.

- Promising Explanation (Olsson et al. (2022)): propose that "induction heads" - attention heads that learn to copy sequences from earlier in the context - are the primary mechanism enabling in-context learning in transformer language models. Through experiments on models of varying sizes, they show that the emergence of induction heads consistently coincides with the rapid improvement of in-context learning ability during training. Furthermore, ablating these heads or modifying the architecture to control their formation directly impacts in-context learning performance. While the evidence is stronger for smaller models, the authors hypothesize that induction heads, defined narrowly in terms of copying random sequences, can be repurposed for more general in-context adaptation. This work provides a promising mechanistic explanation for the important capability of in-context learning in LLMs.

We reference this work to remind us of the underlying neural network architecture that contributes to the ICL capability, which can be a guiding star for us when designing the instruction inputs to LLMs in our approach.

2.1.2 Hallucination in LLMs

Hallucination in LLMs refers to the generation of content that deviates from factual accuracy, posing significant concerns for their application in areas requiring high reliability, such as journalism. Addressing hallucinations is critical to ensure the ethical use of LLMs in evaluating news articles. Here, we summarize related work that explore and analyze the hallucination problem in LLMs.

1. Rawte et al. (2023): This research defines and quantifies the hallucination problem in LLMs, presenting a detailed classification of hallucination types and their severities. It introduces the Hallucination Vulnerability Index (HVI) to rank LLMs based on their susceptibility to hallucinations and proposes mitigation strategies to reduce the incidence of hallucinated content.

2.2 Sourcing in Journalism

Sourcing is a fundamental aspect of journalism, determining the accuracy, credibility, reliability, and ethical quality of news articles. Several studies have explored the importance of sourcing and the challenges associated with evaluating it:

- Coddington and Molyneux (2023): Examine how evidence is represented in news texts from newspapers and digital native outlets between 2007-2019, as journalism has shifted toward more aggregative and intertextual forms. Through a content analysis, they find that first-hand evidence is rarely presented, while non-mediated attributed speech (e.g. interviews) is the most common form of evidence. However, the use of non-mediated speech has declined over time, replaced by increases in mediated speech (e.g. press releases, social media) and

thirdhand evidence (citing other media). Digital outlets relied significantly less on firsthand evidence and non-mediated speech compared to newspapers.

Over the study period, newspapers adopted digital outlets' intertextual approach by citing more social media and other media, while digital outlets reduced their use of thirdhand evidence to mimic newspapers' focus on firsthand and secondhand sourcing. The authors also found that ancillary evidence providing context about sources and evidence-gathering was rare (<25% of sources), requiring audiences to rely on journalists' assumed authority. While national newspapers and digital outlets showed some increases in ancillary evidence over time, the overall presentation of evidence in news texts remains rather opaque. This study highlights important shifts in journalistic sourcing and the need for greater transparency around evidence.

- Gottfried and Liedke (2021): reveals a significant decline in public trust in news media, particularly among Republicans. The percentage of Republicans with at least some trust in national news organizations dropped from 70% in 2016 to 35% in 2021, while Democrats' trust remained relatively stable. This decline has contributed to a growing partisan divide in media trust.

The survey also found a modest decline in trust in local news organizations and consistently low trust in social media as a news source. These findings underscore the challenges faced by journalists in an increasingly polarized and distrustful media landscape.

Given the erosion of public trust in news media, there is a pressing need for innovative solutions to enhance transparency and credibility in journalism. Making the sourcing in news articles more visible through LLMs could be a promising approach to address this issue.

By providing readers with clear and accessible information about the sources used in news articles, LLMs have the potential to foster greater trust and understanding between the public and the media. This increased transparency could help combat the spread of misinformation, encourage more informed public discourse, and ultimately contribute to a healthier democracy.

- Kovach and Rosenstiel (2014): Emphasized the critical role of sourcing in ensuring journalistic integrity and public trust.
- Schudson (2011): Discussed the sociology of news production, highlighting how sourcing practices shape news content and its perceived reliability.

These works establish the context and significance of sourcing in journalism, providing a foundation for our focus on using LLMs to analyze and evaluate sourcing in news articles.

2.3 Automated Evaluation of News Articles

Previous research has attempted to automate various aspects of news analysis, including the evaluation of sourcing. Notable contributions in this area include:

- Fact-Checking Systems (Vlachos and Riedel (2014)): Their work were among the first to introduce the task of automated fact-checking. They defined fact-checking as assigning a truth value to a claim made in a particular context and discussed the construction of a dataset using statements fact-checked by journalists from popular fact-checking websites. The authors also explored potential baseline approaches and challenges in automating the fact-checking process. They proposed

decomposing the task into stages: extracting statements to fact-check, constructing appropriate questions, obtaining answers from relevant sources, and reaching a verdict using these answers.

While the current approach in this thesis focuses on extracting and retrieving sourcing-related entities from news articles rather than directly verifying the claims, the information obtained can potentially be used to support automated fact-checking systems in the future. By accurately identifying sources, their titles, and the statements attributed to them, valuable structured data is provided that could serve as input to fact-checking pipelines. Thus, while this work does not directly involve fact verification, it lays the groundwork for future integration with automated fact-checking systems, potentially enhancing their accuracy and efficiency. This work by Vlachos and Riedel laid the foundation for further research on automated fact-checking and highlighted the potential of NLP techniques to address this task.

- Rapid Source Review (Diakopoulos et al. (2012)): They developed a tool called "Seriously Rapid Source Review" (SRSR) to assist journalists in finding and assessing sources from social media, particularly in the context of breaking news events. The SRSR interface presents a list of potential sources (Twitter users) for a given event, along with various cues and filters to help journalists find and evaluate the credibility of these sources, such as the user's location, aggregate information about the user's network, and the user's historical activity. The system also incorporates machine learning classifiers to categorize users into different types and to identify potential eyewitnesses to the event based on their tweets. Through a user study with professional journalists, the authors found that the eyewitness classifier, user type filters, and visual cues about a source's network and location were particularly useful for

finding and assessing sources.

While SRSR focuses on helping journalists find and evaluate individual sources, this thesis aims to develop methods for extracting and analyzing sourcing information automatically and reliably. Despite these differences in scope and focus, the insights from Diakopoulos et al. (2012)'s study, such as the importance of providing context and multiple cues for source assessment, can inform the design of instructions to LLMs and metrics developed in this thesis.

- Entity Recognition and Attribution (Chiu and Nichols (2016)): Introduced a novel neural network architecture for named entity recognition that combines bidirectional LSTM (BLSTM) with convolutional neural networks (CNN). The BLSTM is used to capture sequential information, while the CNN is employed to extract character-level features. They also proposed a new method of encoding partial lexicon matches in neural networks and compared it to existing approaches. Their system achieved state-of-the-art performance on the CoNLL-2003 and OntoNotes 5.0 datasets, surpassing previous methods that relied heavily on feature engineering and external resources.

While Chiu and Nichols' work focuses on building a specific neural network architecture for named entity recognition, which is a key component in extracting sourcing information, this thesis takes a different approach. Instead of developing a new model for the entity recognition task, we leverage the capability of LLMs to handle named entity recognition.

These studies demonstrate the feasibility and potential benefits of automating the evaluation of news articles, particularly in enhancing transparency and accountability in journalism and journalism ethics.

2.4 Summary

The literature reviewed in this chapter provides a comprehensive overview of the capabilities of LLMs, the importance of sourcing in journalism, and previous efforts to automate the evaluation of news articles. This background informs our approach to using LLMs to analyze and evaluate the sourcing of news articles, addressing both the potential and the challenges involved.

In the following chapters, we will detail our methodology, experiments, and results, contributing to the ongoing efforts to enhance the reliability and transparency of news reporting.

Chapter 3

Our Approach

This chapter outlines our methodology for evaluating the capability of LLMs in extracting sourcing information from news articles. Our approach is designed to provide a comprehensive assessment of LLM performance in this domain, while acknowledging the practical limitations of the thesis. We detail our selection of LLMs, the choice of news articles, the specific sourcing information we aim to extract, our prompting techniques, the creation of our benchmark, and our evaluation methods.

3.1 Selection of Large Language Models

For this thesis, we chose to evaluate two state-of-the-art LLMs:

- Claude 3, developed by Anthropic
- GPT-4, developed by OpenAI

These models were selected due to their advanced capabilities in natural language processing and their widespread use in various applications (OpenAI et al. (2024)). By comparing two different LLMs, we aim to provide insights into the general capabilities of current language models in this domain, rather than focusing on the performance of a single model.

3.2 News Article Dataset

To ensure a diverse and representative test set, we analyzed four news articles covering a range of topics and contexts:

- OpenAI’s Board Fight Reveals a Founder’s Big Weakness
- Border Patrol at California Divide
- Vermont House Overwhelmingly Backs Bill Prohibiting Race-Based Hair Discrimination
- Guaranteed Income Program in California

These articles were selected under the guidance of Subbu Vincent, director of Journalism and Media Ethics at the Markkula Center at Santa Clara University. This expert-guided selection process ensures that our dataset represents a variety of journalistic styles, topics, and sourcing practices, providing a robust testing ground for the LLMs’ capabilities.

It’s important to note that while this dataset is diverse, it is also limited in size due to the resource constraints of this thesis. This limitation will be considered when interpreting our results.

3.3 Sourcing Information Extraction

We focused on extracting five key elements of sourcing information from each article:

- Type of source: Classified as either named or unnamed
- Name of the source: The full name as provided in the article
- Actual attributed statements: Direct quotes or paraphrased statements attributed to the source

- Title of the source: Official position or role of the source, if provided
- Additional characterizations: Any extra information provided by the reporter to justify the source's relevance to the story

This comprehensive set of attributes allows us to assess the LLMs' ability to identify and extract various aspects of sourcing, from basic identification to more nuanced elements like characterizations.

3.4 Prompting Technique

We employed a zero-shot learning approach(Kojima et al. (2023)) for our prompting technique. This choice was made to evaluate the LLMs' base capabilities without additional training or fine-tuning. Our prompting strategy included several key elements:

- Starting with the simplest prompting technique and iteratively refining based on results
- Testing multiple variants of prompts to assess their impact on performance
- Utilizing both system prompts and user prompts to provide context and instructions
- Manually iterating and adjusting prompts based on initial results
- Providing explicit, detailed explanations about the task to the LLMs (Peng et al. (2023))
- Experimenting with different definitions of "source" (e.g., limited to persons, including organizations, or including documents) to assess the impact on extraction performance

- Instructing the LLMs to output results in JSON format for ease of processing

This approach allows us to explore the LLMs' capabilities under various instruction sets and to identify the most effective prompting strategies for this specific task.

3.5 Benchmark Creation: Human Annotation

To create a benchmark for evaluating LLM performance, we constructed a ground-truth dataset with the help of the Markkula Center's Spring 2024 internship class. This class consisted of five undergraduate students from diverse fields, including Communications/Journalism, Political Science, and Engineering. The internship involved a structured process where the students learned and applied journalistic sourcing annotation methods to create a reliable ground-truth dataset. This process was guided by the director of Journalism and Media Ethics to ensure adherence to journalistic standards.

The annotation process involved:

1. Students independently extracting sourcing information from the selected articles
2. Compilation of results in a Google Sheet, later converted to CSV format for evaluation
3. Final review and optimization of results by the director, combining the best elements from each student's work

It's important to note that this human-annotated benchmark may contain some imperfections, such as typos or missing information. These limitations will be considered in our analysis and discussion of results.

3.6 Evaluation Methods

Our evaluation process is designed to provide a comprehensive and nuanced assessment of LLM performance in sourcing extraction. The process involves several steps:

1. **Data Preparation:** Preprocessing both LLM-extracted and human-annotated datasets to ensure consistent formatting and text normalization.
2. **Sentence-Level Matching:** use fuzzy string matching to align sentences from LLM output with human annotations, allowing for minor textual variations.
3. **Source Attribute Comparison:** Comparing extracted source attributes for each matched sentence pair.
4. **Similarity Scoring:** Utilizing similarity scores to quantify match quality for names, titles, and associations, with different thresholds set for each attribute.
5. **Match Determination:** Establishing a "match" based on correct source type identification and meeting similarity thresholds for relevant attributes, with distinct criteria for named and anonymous sources.
6. **Performance Metrics Calculation:** Computing various metrics including LLM Recall, Unique Discovery Rate, and Miss Rate to provide a multi-faceted view of LLM performance.
7. **Cross-Article Analysis:** Repeating the evaluation process across multiple articles to allow for both article-specific and aggregate performance assessment.

This multi-step evaluation process enables us to assess the LLMs' performance in sourcing extraction with a high degree of granularity, considering

both the accuracy of extracted information and the models' ability to identify sources that may have been missed in human annotation.

Chapter 4

Experiments and Results

Discussion

4.1 Experiment Setup

Our experimental process was designed to evaluate the capability of Large Language Models (LLMs) in extracting sourcing information from news articles. The experiment was conducted in two main phases: (1) sourcing information extraction using LLMs, and (2) evaluation of the LLM outputs against human-annotated ground truth data. We utilized two state-of-the-art LLMs: GPT-4 and Claude 3.

4.1.1 Data Preparation

We selected four diverse news articles for our analysis, as described in Chapter 3. These articles were saved as plain text (.txt) files to facilitate processing. The articles covered a range of topics and contexts, ensuring a robust test of the LLMs' capabilities across different journalistic styles and subject matters.

4.1.2 LLM Analysis Process

For the LLM analysis, we developed a Python script that performs the following steps:

1. API Setup: The script initializes API clients for both OpenAI (for GPT-4) and Anthropic (for Claude 3) using secure API key management.
2. Prompt Loading: The latest versions of system and user prompts are dynamically loaded from a designated directory. This approach allows for easy iteration and testing of different prompt strategies.
3. Article Processing: Each article is processed sequentially through both GPT-4 and Claude 3.
4. LLM Instructing: The script provides each LLM with system and user prompts, along with the article text, instructing the model to perform the sourcing extraction task.
5. Output Parsing: The LLM responses, expected to be in JSON format, are extracted and parsed.
6. Data Saving: The results are saved in both JSON and CSV formats for each LLM and article combination. The CSV format includes columns for Source Type, Name, Title, Association, and Sourced Statement.
7. Experiment Tracking: Each run of the experiment is saved in a timestamped directory, including the prompts used, to ensure reproducibility.

4.1.3 LLM Prompt Strategies

Our experiment utilized two distinct sets of prompts to evaluate the LLMs' performance under different instruction scenarios:

V0: Person-Only Prompts

- System Prompt: Instructed the LLM to identify only individual persons as sources.
- User Prompt: Requested extraction of named and anonymous sources, including names, titles, associations, and sourced statements.

- Key Feature: Focused solely on human sources, excluding organizations.

V1: Person and Organization Prompts

- System Prompt: Expanded the definition of sources to include both persons and organizations.
- User Prompt: Similar to v0, but with the addition of handling organizational sources (e.g., using 'null' for titles of organizational sources).
- Key Feature: Broadened the scope to include organizations as potential named sources.

Both prompt sets emphasized the importance of capturing multiple sourced statements per source and provided detailed definitions of key terms such as "Source," "Sourced Statements," "Title," and "Associations to the story."

The use of these two prompt strategies allowed us to assess:

- The LLMs' ability to adapt to different definitions of sources.
- The impact of including or excluding organizational sources on overall performance.
- The models' flexibility in handling nuanced instructions about source types and attribution.

The complete prompts used in our experiments are provided in Appendix A.

4.1.4 Evaluation Process

The evaluation of the LLM outputs is performed using another Python script, which implements the following steps:

1. Data Loading: The script loads the human-annotated ground truth data and the LLM-generated data for each article.

2. Text Preprocessing: All text data is cleaned and normalized, including lowercasing, whitespace removal, and sentence tokenization.

3. Sentence Matching: The script uses fuzzy string matching to align sentences from the LLM output with those in the human annotations, allowing for minor variations in text.

4. Attribute Comparison: For each matched sentence pair, the script compares source attributes (Type, Name, Title, Association) between the ground truth and LLM output.

5. Performance Metrics Calculation: The script calculates several performance metrics:

- LLM Recall: The proportion of ground truth sources correctly identified by the LLM.

$$\text{LLM Recall} = \frac{\text{Match Counts} + \text{LLM Unique Discover Counts}}{\text{Total Unique Sentences}} \quad (4.1)$$

Match Counts: Number of sentences that match in source type and either name (for named sources) or title/association (for anonymous sources).

LLM Unique Discover Counts: Number of sentences found by LLM but not in ground truth.

Total Unique Sentences: Total number of unique sentences across ground truth and LLM output.

Explanation: This metric measures the LLM's ability to identify and correctly attribute sourced statements, including both those in the ground truth and any additional valid discoveries.

- Unique Discovery Rate: The proportion of sources identified by the LLM that were not in the ground truth.

$$\text{LLM Unique Discover Rate} = \frac{\text{LLM Unique Discover Counts}}{\text{Total Unique Sentences}} \quad (4.2)$$

Explanation: This metric measures the proportion of unique sourced statements discovered by the LLM that were not in the ground truth.

- Miss Rate: The proportion of ground truth sources not identified by the LLM.

$$\text{LLM Miss Rate} = \frac{\text{Human Unique Discover Counts}}{\text{Total Unique Sentences}} \quad (4.3)$$

Human Unique Discover Counts: Number of sentences in ground truth not found by LLM.

Explanation: This metric measures the proportion of sourced statements in the ground truth that the LLM failed to identify or attribute correctly.

- Match Rates for Name, Title, Association, and Source Type: The proportion of correctly identified attributes for matched sources.

$$\text{Name Match Rate} = \frac{\text{Count of sentences where Name Match Score} > 80}{\text{Total sentences found by both human and LLM}} \quad (4.4)$$

Explanation: This metric measures the proportion of sentences where the names of sources match closely between the ground truth and LLM output. A score above 80 is considered a match.

$$\text{Title Match Rate} = \frac{\text{Count of sentences where Title Match Score} > 55}{\text{Total sentences found by both human and LLM}} \quad (4.5)$$

Explanation: This metric measures the proportion of sentences where the titles of sources match reasonably well between the ground truth and LLM output. A score above 55 is considered a match.

$$\text{Association Match Rate} = \frac{\text{Count of sentences where Association Match Score} > 55}{\text{Total sentences found by both human and LLM}} \quad (4.6)$$

$$\text{Source Type Match Rate} = \frac{\text{Count of sentences where Source Type Match is 'Yes'}}{\text{Total sentences found by both human and LLM}} \quad (4.7)$$

Explanation: This metric measures the proportion of sentences where the source type (e.g., named or anonymous) matches exactly between the ground truth and LLM output.

6. Results Compilation: The script compiles the results for all articles and both LLMs into a single performance table.

7. Visualization: The script generates various plots to visualize the performance comparisons between the LLMs across different metrics and articles.

4.1.5 Experimental Parameters

LLM Models:

- GPT-4 (version: gpt-4-turbo)
- Claude 3 (version: claude-3-opus-20240229)

LLM Configuration:

- Temperature: 0 (to maximize deterministic and focused outputs)
- Maximum response length: 4096 tokens (the maximum allowed by the API)

Similarity Scores Computation:

The similarity scores in this script are computed using the *Levenshtein distance ratio*, implemented through the *fuzz.ratio()* function from the *fuzzywuzzy* library.

Computation: The similarity score is calculated as:

$$\text{Similarity Score} = \frac{\text{Levenshtein Distance}}{\text{Length of Longer String}} \times 100 \quad (4.8)$$

Where Levenshtein Distance is the minimum number of single-character edits (insertions, deletions, or substitutions) required to change one string into another.

The score ranges from 0 to 100. 0 means the strings are completely different. 100 means the strings are identical. Higher scores indicate greater similarity.

Thresholds used in the evaluation:

- Sentence Match: > 60 similarity. We set the threshold slightly above 50 to allow for minor variations in text, so that it can handle possible different parsing of sentences from LLM extrated data and human annotated data.
- Name Match: > 80 similarity. We set a higher bar for accuracy in identifying names.
- Title and Association Match: > 55 similarity. Titles and associations of sources in news articles often have more variability in how they are expressed compared to names. A lower threshold allows for these variations while still capturing the essence of the role.

Output Format:

Structured JSON with fields for Source Type, Name, Title, Association, and Sourced Statements.

4.1.6 Experimental Runs

To ensure the robustness of our findings and to assess the consistency of LLM outputs, we conducted multiple experimental runs:

- Each set of prompts was used in three separate experimental runs.
- This resulted in a total of six runs per article (3 runs \times 2 prompt sets).
- The experiment was conducted on all four selected news articles.

This repeated testing approach allows us to:

- Evaluate the consistency of LLM outputs across multiple runs with the same prompts.
- Identify any variability in the models' performance or output.
- Increase the reliability of our findings by basing them on multiple data points rather than a single run.
- Assess whether any observed differences between prompt strategies are consistent across repeated trials.

By running each experiment multiple times, we can provide a more comprehensive analysis of the LLMs' capabilities and limitations in the task of sourcing extraction from news articles.

4.2 Results Discussion

After conducting three experimental runs using prompt version 0 (person-only sources), we obtained a set of performance metrics for both Claude and GPT-4 across the four articles. This section presents an analysis of these results, focusing on the consistency across runs, performance comparisons between the models, and article-specific observations.

4.2.1 Consistency Across Runs

Claude’s Performance

Claude demonstrated high consistency across the three runs for most articles and metrics:

TABLE 4.1: Claude’s LLM Recall Consistency Across Three Runs

Article	Run 1	Run 2	Run 3
vermont_bill	0.914	0.941	0.941
ca_guaranteed_income	0.467	0.450	0.450
openai_board	0.613	0.618	0.618
border_patrol_towers	0.256	0.279	0.279

GPT-4’s Performance

GPT-4 showed more variability in its performance across the three runs:

TABLE 4.2: GPT-4’s LLM Recall Consistency Across Three Runs

Article	Run 1	Run 2	Run 3
vermont_bill	0.892	0.944	0.343
ca_guaranteed_income	0.433	0.484	0.753
openai_board	0.545	0.290	0.303
border_patrol_towers	0.146	0.146	0.122

4.2.2 Performance Comparison

LLM Recall

Average LLM Recall across the three runs:

TABLE 4.3: Average LLM Recall Comparison Between Claude and GPT-4

Article	Claude	GPT-4
vermont_bill	0.932	0.726
ca_guaranteed_income	0.456	0.557
openai_board	0.616	0.379
border_patrol_towers	0.271	0.138

- Claude generally outperformed GPT-4 in terms of LLM Recall, especially in the "border_patrol_towers" and "openai_board" articles.
- However, GPT-4 showed superior performance in one run of the "ca_guaranteed_income" article (0.753 vs 0.45).
- This suggests that while Claude may be more consistent, GPT-4 has the potential for higher peak performance in some cases.

Unique Discovery Rate

Average Unique Discovery Rate across the three runs:

TABLE 4.4: Average Unique Discovery Rate Comparison Between Claude and GPT-4

Article	Claude	GPT-4
vermont_bill	0.155	0.194
ca_guaranteed_income	0.000	0.129
openai_board	0.180	0.164
border_patrol_towers	0.047	0.000

- Both models showed relatively low Unique Discovery Rates across all articles.

- GPT-4 had a notably high Unique Discovery Rate (0.355) in one run of the "ca_guaranteed_income" article.
- The generally low Unique Discovery Rates suggest that both models are more likely to identify sources already present in the ground truth rather than finding additional sources.

Name and Source Type Match Rates

Average Name Match Rate and Source Type Match Rate across the three runs:

TABLE 4.5: Average Name Match Rate and Source Type Match Rate

Article	Model	Name Match Rate	Source Type Match Rate
vermont_bill	Claude	1.000	1.000
	GPT-4	1.000	1.000
ca_guaranteed_income	Claude	1.000	1.000
	GPT-4	1.000	1.000
openai_board	Claude	0.921	0.861
	GPT-4	0.863	0.819
border_patrol_towers	Claude	0.897	0.976
	GPT-4	0.849	1.000

- Both models consistently performed well in Name Match Rate and Source Type Match Rate across all articles and runs.
- This high performance indicates that when sources are correctly identified, both models are accurate in determining the source type and extracting the correct name.

Title and Association Match Rates

Average Title Match Rate and Association Match Rate across the three runs:

- Performance in these metrics was more variable, with both models showing inconsistencies across different articles and runs.

TABLE 4.6: Average Title Match Rate and Association Match Rate

Article	Model	Title Match Rate	Association Match Rate
vermont_bill	Claude	1.000	0.300
	GPT-4	1.000	0.247
ca_guaranteed_income	Claude	0.561	0.707
	GPT-4	0.000	0.590
openai_board	Claude	0.843	0.384
	GPT-4	0.360	0.156
border_patrol_towers	Claude	0.818	0.135
	GPT-4	0.849	0.095

- This variability suggests that extracting and matching titles and associations is a more challenging task for both models compared to identifying names and source types.

4.2.3 Article-Specific Observations

When analyzing the performance of the LLMs across different articles, it's crucial to consider the types of sources present in each article. The varying composition of source types may explain some of the performance differences we observed.

"vermont_bill":

- Source types: Primarily named individuals
- Performance: Both models performed well on this article, with Claude showing more consistency.
- Analysis: The high performance on this article likely stems from the straightforward nature of the sources. Named individuals are typically easier for LLMs to identify and extract information about, which aligns with the strong performance we observed.

"ca_guaranteed_income":

- Source types: Named individuals, Unnamed groups of people (e.g., "Researchers and advocates"), Named organizations (e.g., Stanford Guaranteed Income Lab)
- Performance: This article saw the most variability in performance, especially for GPT-4.
- Analysis: The mix of source types in this article, particularly the inclusion of unnamed groups and organizations, may explain the inconsistent performance. LLMs might struggle more with identifying and categorizing unnamed groups or distinguishing between individual and organizational sources.

"openai_board":

- Source types: Named individuals, Anonymous individuals, Named organizations (e.g., OpenAI)
- Performance: Claude consistently outperformed GPT-4 in LLM Recall for this article.
- Analysis: The presence of anonymous sources adds complexity to the task. Claude's superior performance suggests it might be better at handling a mix of named and anonymous sources, as well as distinguishing between individual and organizational sources in this context.

"border_patrol_towers":

- Source types: Named individuals, Unnamed groups of people (e.g., "Advocates"), Organizations (e.g., Dept of Homeland Security)
- Performance: Both models struggled with this article, showing lower LLM Recall compared to other articles.

TABLE 4.7: Comparison of Average LLM Recall Between Prompt Versions 0 and 1

Article	Model	v0 Avg LLM Recall	v1 Avg LLM Recall	Difference (v1 - v0)
vermont_ bill	CLAUDE	0.932	0.611	-0.321
	GPT4	0.726	0.350	-0.376
ca_guaranteed_ income	CLAUDE	0.456	0.367	-0.089
	GPT4	0.557	0.652	+0.095
openai_ board	CLAUDE	0.616	0.469	-0.147
	GPT4	0.379	0.324	-0.055
border_patrol_ towers	CLAUDE	0.271	0.318	+0.047
	GPT4	0.138	0.154	+0.016

- Analysis: The diverse mix of source types, including unnamed groups and organizations, likely contributed to the lower performance. This article seems to present the most complex sourcing structure among the four, which explains why both models found it challenging.

4.2.4 Comparison of Prompt Versions

We conducted experiments using two different prompt versions: version 0 (v0) focused on individual sources, and version 1 (v1) which expanded the definition to include organizations as potential sources. By comparing the results from these two prompt versions, we can gain insights into how prompt design affects LLM performance in sourcing extraction tasks.

Overall Performance Comparison

To compare the overall performance between the two prompt versions, we'll look at the average LLM Recall across all three runs for each article and model as shown in Table 4.7

Key Observations

Performance Variability:

- Claude generally showed more consistent performance across runs in both v0 and v1.
- GPT4 continued to show more variability, especially in the "ca_guaranteed_income" article.

Impact on Different Articles:

- "vermont_bill": Both models saw a significant decrease in performance with v1.
- "ca_guaranteed_income": Claude's performance slightly decreased, while GPT4's improved notably.
- "openai_board": Both models saw a decrease in performance, with Claude affected more significantly.
- "border_patrol_towers": Both models saw a slight improvement in performance with v1.

Source Type Handling: The inclusion of organizations as potential sources in v1 seemed to have a mixed impact, improving performance in some cases (e.g., "border_patrol_towers") but decreasing it in others (e.g., "vermont_bill").

Model-Specific Trends:

- Claude's performance generally decreased with v1, except for a slight improvement in "border_patrol_towers".
- GPT4 showed more varied results, with significant improvement in "ca_guaranteed_income" but decreased performance in "vermont_bill" and "openai_board".

Analysis of Prompt Version Impact

Expanded Source Definition: The inclusion of organizations as potential sources in v1 seems to have had a complex impact on performance. While it potentially improved the models' ability to handle organizational sources, it may have also introduced confusion in articles predominantly featuring individual sources.

Trade-offs in Performance: The decrease in performance for "vermont_bill" (which primarily features individual sources) suggests that broadening the source definition may have come at the cost of reduced accuracy for individual source identification.

Adaptation to Complex Source Structures: The improved performance in "border_patrol_towers" and GPT4's improvement in "ca_guaranteed_income" indicate that v1 may be better suited for articles with a mix of individual and organizational sources.

Model-Specific Responses to Prompt Changes: The differing responses of Claude and GPT4 to the prompt change suggest that the two models may have different strengths and weaknesses in adapting to more complex source definitions.

Consistency vs. Adaptability: Claude's more consistent performance across both prompt versions suggests it may be more robust to prompt changes, while GPT4's variable performance indicates it might be more sensitive to prompt modifications, sometimes leading to significant improvements.

Chapter 5

Conclusion and Future Work

5.1 Limitations

Before presenting our conclusions, it's important to acknowledge the limitations of this thesis:

Narrow Source Definition: Our focus was limited to persons and organizations as sources, excluding other important source types like documents, and common knowledge. This narrow scope restricts the broader applicability of our findings.

Data Processing Constraints: We used raw text versions of articles, omitting potentially valuable contextual information from links, images, and other media. The inclusion of extraneous data like advertisements may have impacted our results' accuracy.

Limited Dataset: Our thesis analyzed only four news articles, with ground truth established through limited student annotations and expert supervision. This small sample size may not fully represent the diverse landscape of journalistic sourcing practices.

Limitations of Levenshtein Distance Ratio: Our use of the Levenshtein distance ratio for similarity comparisons was chosen to ensure LLMs extract the exact phrasing used by journalists, which is crucial for direct quotes and

attributions. While effective for aligning sentences and detecting minor variations, this method falls short when comparing nuanced elements like "association to the story". The Levenshtein ratio fails to capture semantic similarities, potentially missing meaning-equivalent but lexically different phrases. This limitation affects the accuracy of our association matching scores and suggests the need for a dual approach in future research: maintaining lexical comparison for direct extractions while incorporating semantic similarity measures for contextual information. Such an approach would better balance the need for exact quotation with the understanding of broader contextual meanings in journalistic sourcing.

5.2 Conclusion

This thesis set out to explore the potential of Large Language Models (LLMs) in automating the extraction and analysis of sourcing information from news articles. Our research focused on evaluating the ability of two state-of-the-art LLMs, GPT-4 and Claude 3, to identify and categorize various source types across four diverse news articles.

Key findings include:

- **Promising Potential:** Both LLMs demonstrated capabilities in extracting sourcing information, particularly in identifying named sources and their associated statements. This suggests that LLMs could play a valuable role in developing tools for automated sourcing analysis in journalism.
- **Performance Variability:** We observed significant variations in performance across different articles and between the two LLMs. This highlights the complexity of the task and the need for robust, adaptive solutions.

- **Prompt Sensitivity:** Our experiments with different prompt versions revealed that LLMs' performance can be significantly influenced by the specificity and framing of instructions. This underscores the importance of careful prompt engineering in leveraging LLMs for specialized tasks.
- **Consistency vs. Adaptability:** Claude generally showed more consistent performance across different prompt versions, while GPT-4 exhibited greater variability but also potential for higher peak performance in some cases.
- **Challenges with Complex Sources:** Both models struggled with more nuanced sourcing structures, such as unnamed groups or organizations as sources. This indicates areas for improvement in LLM capabilities or in the design of prompts for handling diverse source types.

Our findings suggest that while LLMs show promise in automating aspects of sourcing analysis in journalism, there remain significant challenges to overcome before they can be reliably deployed in real-world applications.

5.3 Future Work

Building on our findings, we propose several directions for future research:

5.3.1 Enhanced Instruction Design

Develop more sophisticated prompting techniques, potentially incorporating feedback mechanisms or advanced prompt engineering strategies to improve LLM performance on complex sourcing structures.

5.3.2 Larger and More Diverse Datasets

Expanding the ground-truth dataset to include a wider variety of news articles from different sources and regions will enhance the robustness and generalizability of the evaluation. This would also involve incorporating multiple annotators to ensure consistency and reliability in the annotations.

5.3.3 Real-Time Applications

Developing real-time applications for automated sourcing evaluation in journalism can provide immediate feedback to journalists and editors, promoting better sourcing practices. This could include integrating LLMs into content management systems or news verification platforms.

5.3.4 On-demand Source Literacy tools for Everyday News Consumers

The findings of this thesis lay the groundwork for the development of on-demand source literacy tools designed specifically for everyday news consumers. As the media landscape continues to evolve and the amount of information available online grows exponentially, it becomes increasingly challenging for individuals to assess the credibility and reliability of news sources. By leveraging the capabilities of LLMs in extracting and analyzing sourcing information, future research could focus on creating user-friendly applications that provide real-time insights into the sources cited in news articles. These tools could offer a range of features, such as highlighting the types of sources used, identifying potential biases or conflicts of interest, and providing context about the sources' expertise or relevance to the story. Additionally, the tools could incorporate educational components to help users better understand the importance of source evaluation and how to critically

assess the information they consume. By empowering everyday news consumers with accessible and intuitive source literacy tools, we can promote a more informed and discerning public, ultimately contributing to a healthier and more trustworthy media ecosystem.

5.3.5 Addressing Ethical Considerations

As LLMs become more integrated into journalistic practices, it is essential to address ethical considerations, including the transparency of automated systems, the potential for bias in model outputs, and the implications for journalistic integrity. Future research should focus on developing ethical guidelines and frameworks for the use of AI in journalism.

5.3.6 Cross-Disciplinary Collaboration

Further collaboration between technologists, journalists, and media scholars will be crucial in advancing the application of LLMs in journalism. Such interdisciplinary efforts can lead to more holistic solutions that address both technical and practical challenges in news sourcing evaluation.

5.4 Final Thoughts

This thesis represents a crucial step towards harnessing the power of AI to enhance transparency and accountability in journalism. By demonstrating both the potential and limitations of current LLM technology in sourcing analysis, we have laid the groundwork for future innovations in this critical area.

As we continue to refine these technologies, it's essential to maintain a balance between leveraging AI capabilities and preserving the fundamental principles of journalistic integrity. The ultimate goal is not to replace human

judgment in journalism, but to augment it, providing tools that can enhance the quality, reliability, and transparency of news reporting in our increasingly complex information landscape.

The journey of integrating AI into journalistic practices is just beginning. By addressing the challenges identified in this research and pursuing the proposed avenues for future work, we can work towards a future where AI serves as a powerful ally in upholding the highest standards of journalism, fostering an informed and discerning public in the digital age.

References

- Anthropic (2024). Introducing the next generation of claude.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Nee-lakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.
- Chiu, J. P. and Nichols, E. (2016). Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370.
- Coddington, M. and Molyneux, L. (2023). Making sources visible: Representation of evidence in news texts, 2007–2019. *Journalism Practice*, 17(4):664–682.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.
- Diakopoulos, N., De Choudhury, M., and Naaman, M. (2012). Finding and assessing social media information sources in the context of journalism. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 2451–2460, New York, NY, USA. Association for Computing Machinery.
- Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., Li, L., and Sui, Z. (2023). A survey on in-context learning.
- Gao, L., Chaudhary, A., Srinivasan, K., Hashimoto, K., Raman, K., and Bendersky, M. (2024). Ambiguity-aware in-context learning with large language models.

- Goel, A., Gueta, A., Gilon, O., Liu, C., Erell, S., Nguyen, L. H., Hao, X., Jaber, B., Reddy, S., Kartha, R., Steiner, J., Laish, I., and Feder, A. (2023). Llms accelerate annotation for medical information extraction. In Hegselmann, S., Parziale, A., Shanmugam, D., Tang, S., Asiedu, M. N., Chang, S., Hartvigsen, T., and Singh, H., editors, *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pages 82–100. PMLR.
- Gottfried, J. and Liedke, J. (2021). Partisan divides in media trust widen, driven by a decline among republicans. *Pew Research Center*.
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2021). Measuring massive multitask language understanding.
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2023). Large language models are zero-shot reasoners.
- Kovach, B. and Rosenstiel, T. (2014). *The Elements of Journalism: What Newspeople Should Know and the Public Should Expect*. Three Rivers Press, New York, 3rd edition.
- Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. (2022). Rethinking the role of demonstrations: What makes in-context learning work?
- Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J.,

Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. (2022). In-context learning and induction heads. *Transformer Circuits Thread*. <https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html>.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kopic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey,

C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Nee-lakantan, A., Ngo, R., Noh, H., Ouyang, L., O’Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sas-try, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. (2024). Gpt-4 technical report.

Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., Huang, Q., Liden, L., Yu, Z., Chen, W., and Gao, J. (2023). Check your facts and try again: Improving large language models with external knowledge and automated feedback.

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., and Yang, D. (2023). Is chatgpt a general-purpose natural language processing task solver?

- Rawte, V., Chakraborty, S., Pathak, A., Sarkar, A., Tonmoy, S. T. I., Chadha, A., Sheth, A., and Das, A. (2023). The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2541–2573, Singapore. Association for Computational Linguistics.
- Reiss, M. V. (2023). Testing the reliability of chatgpt for text annotation and classification: A cautionary remark.
- Schudson, M. (2011). *The Sociology of News*. W. W. Norton Company, New York, 2nd edition.
- Singh, C., Inala, J. P., Galley, M., Caruana, R., and Gao, J. (2024). Rethinking interpretability in the era of large language models.
- Törnberg, P. (2023). Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning.
- Vlachos, A. and Riedel, S. (2014). Fact checking: Task definition and dataset construction. In Danescu-Niculescu-Mizil, C., Eisenstein, J., McKeown, K., and Smith, N. A., editors, *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.
- Wang, X., Zhu, W., Saxon, M., Steyvers, M., and Wang, W. Y. (2024). Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning.
- Zhang, Y. (2024). Unveiling llm mechanisms through neural odes and control theory.

Appendix A

Appendix A

System Prompt - V0 - Persons Only

You are an AI assistant that analyzes news articles to identify all Named and Anonymous Sources, their Titles, Associations to the story, and their related Sourced Statements.(Note: There may be multiple sourced statements from one source. Find all of them.) Below is how you define the terminology:

Source: A person that the reporter is citing or quoting or paraphrasing to attribute claims or viewpoints or facts or experiences in the story. Sources may be named or anonymous.

Sourced Statements: Statements that include viewpoints, experiences, criticism, questioning, support or other expressions from a source, captured in direct quotes or indirect speech.Remember Sourced Statements not only include the speech of the source, i.e. what the source said(direct quotes, indirect speech, and paraphrasing) but also include all other forms of attribution that describe the source's conduct or position or attitude, for e.g. lines that state that a source criticized, or supported, or questioned, or decided something etc. In other words, all statements that would not have been present in the story if the reporter had not spoken to or contacted the source.

Title: The occupation or position or designation held by the person in an organization or in an independent capacity. If the person is a legislator, the title includes the constituency. (Note: for anonymous sources, the title may

not be provided, and only the association to the story is mentioned. In this case, put a 'null' as a Title.)

Associations to the story: Characterizations or descriptions of the source that the reporter has used to justify why the source should be present in the story.

User Prompt - V0

Please analyze the following news article and provide the requested information in pure JSON format:{article_text} Extract the following details: 1. Named sources (use 'NamedSources' as the key):

- - Names (use 'Name' as the key)
- - Titles (use 'Title' as the key)
- - Associations to the story (use 'Association' as the key)
- - Sourced Statements (use 'SourcedStatement' as the key)

2. Anonymous sources (use 'AnonymousSources' as the key):

- - Titles (use 'Title' as the key)
- - Associations to the story (use 'Association' as the key)
- - Sourced Statements (use 'SourcedStatement' as the key)

System Prompt - V1 - Persons and Organizations

You are an AI assistant that analyzes news articles to identify all Named and Anonymous Sources, their Titles (if applicable), Associations to the story, and their related Sourced Statements. (Note: There may be multiple sourced

statements from one source. Find all of them.)Below is how you define the terminology:

Source: A person or organization that the reporter is citing, quoting, or paraphrasing to attribute claims, viewpoints, facts, or experiences in the story. Sources may be named or anonymous. Organizations can be considered as Named sources.

Sourced Statements: Statements that include viewpoints, experiences, criticism, questioning, support or other expressions from a source, captured in direct quotes or indirect speech. Remember Sourced Statements not only include the speech of the source (i.e., what the source said in direct quotes, indirect speech, and paraphrasing) but also include all other forms of attribution that describe the source's conduct, position, or attitude. For example, lines that state that a source criticized, supported, questioned, or decided something, etc. In other words, all statements that would not have been present in the story if the reporter had not spoken to or contacted the source.

Title: For individual sources, this is the occupation, position, or designation held by the person in an organization or in an independent capacity. If the person is a legislator, the title includes the constituency. For organizational sources, no title is required. (Note: for anonymous sources, the title may not be provided, and only the association to the story is mentioned. In this case, put a 'null' as a Title.)

Associations to the story: Characterizations or descriptions of the source that the reporter has used to justify why the source should be present in the story.

User Prompt - V1

Please analyze the following news article and provide the requested information in pure JSON format:
`article_text`
Extract the following details : 1. Named sources (use 'NamedSource'

Names (use 'Name' as the key). This can be a person's name or an organization's name

Titles (use 'Title' as the key). For individual sources only; use 'null' for organizational sources.

Associations to the story (use 'Association' as the key)

Sourced Statements (use 'SourcedStatement' as the key)

2. Anonymous sources (use 'AnonymousSources' as the key):

- Titles (use 'Title' as the key)
- Associations to the story (use 'Association' as the key)
- Sourced Statements (use 'SourcedStatement' as the key)