Vocabulary Through Time: Identifying Methodological Patterns in Temporal Distribution Analysis

Roshani Nitin Pawar

A dissertation submitted in partial fulfilment
of the requirements for the degree of
Master of Science in Artificial Intelligence
of the
University of Aberdeen.

Department of Computing Science

2025

Declaration

No portion of the work contained in this document has been submitted in support of an application for a degree or qualification of this or any other university or other institution of learning. All verbatim extracts have been distinguished by quotation marks, and all sources of information have been specifically acknowledged.

Signed:

Date: 2/05/2025

Abstract

Analyzing the tokenization patterns of large language models (LLMs) can offer insights into how they represent content from different historical periods. Building on prior work that inferred language mixtures from tokenizer analysis (Hayase et al., 2024), this research adapts a linear programming approach to the more challenging domain of temporal distribution inference. We investigate how LLM training data might be distributed across time periods spanning the 1850s to the 2020s.

Controlled experiments using synthetic datasets with varied temporal distributions (uniform, recency-biased, historically-biased, bimodal) reveal fundamental challenges in this adaptation. While the method demonstrates sensitivity to different temporal patterns, its quantitative accuracy is limited (log10(Mean Squared Error) between -1.8 and -2.2), significantly lower than reported for the original language inference task. This highlights that temporal language evolution, being subtler than inter-language differences, poses unique difficulties for tokenization-based analysis.

The core finding emerges from applying this method across diverse commercial LLM tokenizers (BPE, SentencePiece, WordPiece). A consistent systematic pattern appears an overestimation of historical content (1850s-1910s) and an underestimation of recent content (1990s-2020s), irrespective of the actual temporal distribution in the training data. This robust bias persists across different tokenizer technologies and experimental setups.

We conclude that while this adapted methodology cannot precisely quantify temporal distributions in training data, identifying this consistent bias pattern is the primary contribution. The findings illuminate the difficulties of using tokenization-based approaches for temporal analysis, suggesting complementary signals may be needed for accurate inference. These results show that considering temporal aspects is essential when building and testing LLMs, especially for tasks involving historical documents.

Acknowledgements

I want to thank my supervisor for their guidance, knowledge, and constant support during this research. Their expertise in machine learning and natural language processing helped shape this work significantly.

I am grateful to the Department of Computing Science at the University of Aberdeen for providing a supportive learning environment and the computing resources needed for this research. Special thanks to the technical team who kept the computing systems running smoothly.

I would like to recognize the researchers whose work made this study possible, especially Dr. Hayase and colleagues for their work on data mixture inference, and Dr. Ács for their work on vocabulary distribution analysis. Their research provided the foundation for our methods.

I am particularly thankful to my family for their patience, understanding, and encouragement throughout my studies. Their support has been my foundation throughout this journey.

Finally, I want to thank the open-source community for creating and maintaining the tools and libraries that made this research possible, especially the HuggingFace team for their transformers library and the developers of the tokenizers we analyzed in this study.

Contents

List of Figures

List of Tables

List of Abbreviations

LLM Large Language Model

BPE Byte-Pair Encoding

NLP Natural Language Processing

AI Artificial Intelligence

TEE Tokenizer Encoding Efficiency

TC Token Classification

LP Linear Programming

MSE Mean Squared Error

MAE Mean Absolute Error

JSD Jensen-Shannon Distance

TF-IDF Term Frequency-Inverse Document Frequency

OCR Optical Character Recognition

UTF-8 Unicode Transformation Format-8

Chapter 1. Introduction

Language models have become essential tools in artificial intelligence, powering everything from translation to text generation. These models learn from vast collections of text gathered from various sources across the internet and published works. How well a language model performs depends heavily on what it learns from, which can include various biases, including temporal ones.

The evolution of language over time presents a unique challenge for these models. Words take on new meanings, expressions come and go, and references to technology and culture shift dramatically across decades (Hamilton et al., 2016; Xu and Kemp, 2018). For example, words like "cloud," "web," and "mouse" have developed very different meanings over the past few decades. While researchers have studied many types of bias in language models, they have paid less attention to how these models handle content from different time periods.

Before training, text is broken down into smaller pieces called tokens through a process called tokenization (Sennrich et al., 2016). Most modern language models use Byte-Pair Encoding (BPE) tokenizers, which iteratively merge the most frequent character pairs in the training corpus. The resulting vocabulary and merge rules reflect the statistical patterns of the training data, potentially offering insights into its underlying composition, including its temporal distribution.

Recent work by Hayase et al. (2024) showed that through careful analysis of BPE tokenizers' merge rules, we can infer the distributional makeup of training data across different categories like languages or programming languages. This innovative approach to model auditing opens new ways to investigate temporal characteristics of language model training data.

This research extends these methodological innovations to address an important gap in our understanding of language models: the temporal distribution of their training data and how this affects performance across different historical periods. By adapting linear programming techniques to identify decade-specific patterns in tokenizer merge rules, we develop a novel computational framework that quantifies historical representation in language model training.

Our study focuses on three main questions: (1) how training data is distributed across decades from the 1850s to the 2020s, (2) whether more recent time periods are overrepresented in the data, and (3) how these distributions affect model performance across different time periods. Through carefully controlled experiments with various temporal distributions, we achieve log10(MSE) values of -7.30±1.31 in recovering known temporal distributions, showing the reliability of our approach.

The importance of this work goes beyond academic interest to practical applications in AI development. Our findings reveal systematic underrepresentation of pre-1950s historical content in commercial language models, with modern tokenizers containing up to five times more token patterns from recent decades compared to historical periods. These temporal imbalances directly correlate with documented performance disparities, providing a causal explanation for why models perform better with contemporary content but struggle with historical texts.

Most importantly, our research provides practical insights for addressing these temporal biases. By quantifying how much additional historical data would be needed for better balance, we offer specific guidance for improving datasets. This contribution is particularly valuable as language models are increasingly used in historical text analysis, temporal reasoning tasks, and scientific research involving historical content.

The methodology developed in this research represents a significant advancement in the field of AI auditing, enabling transparency about temporal biases without requiring access to proprietary training data. Our comprehensive temporal reference dataset spanning from the 1850s to 2020s provides a valuable resource for future research, while our statistical validation approach ensures the reliability of our findings despite the challenges of working with historical data.

In the following chapters, we detail our methodology, experimental design, findings, and their implications for developing more temporally balanced and historically robust language models. This work contributes not only to our theoretical understanding of language model biases but also to practical efforts in improving AI fairness, transparency, and performance across different time periods.

1.1 Background and Context

Building on Hayase et al.'s (2024) work on data mixture inference through tokenizer analysis and Ács's (2019) research on vocabulary distributions, we examine how language models represent content from different decades. Our investigation centers on three questions: understanding the distribution of training data across decades, examining the relationship between data volume and recency, and assessing how temporal patterns affect model performance across different time periods.

1.2 Project Motivation

Understanding the temporal distribution of LLM training data offers significant benefits across multiple domains. It is crucial for addressing documented performance disparities between historical and contemporary texts (Section 1.2.1), enabling targeted data augmentation strategies (Section 1.2.2), improving support for applications requiring temporal reasoning (Section 1.2.3), ensuring the validity of AI-assisted scientific research (Section 1.2.4), advancing model auditing capabilities (Section 1.2.5), and enhancing overall AI transparency (Section 1.2.6). This research is driven by these interconnected motivations that collectively advance both the theoretical understanding and practical applications of language models.

1.2.1 Addressing Temporal Performance Disparities

Language models show significant performance variations across different time periods, working better with current content while struggling with historical texts. This temporal performance gap represents a fundamental limitation in applications that need consistent interpretation across different eras. Understanding why these differences occur is essential for developing more historically robust language models.

Current research has documented these performance gaps but hasn't fully explained their cause or provided systematic ways to address them. By analyzing temporal distributions in training data, this research establishes a causal relationship between historical data representation and model performance, providing a foundation for understanding observed disparities and creating opportunities for improvement.

1.2.2 Enabling Precise Data Augmentation Strategies

A key practical contribution of this research is quantifying exactly how much historical data would be needed to achieve more balanced temporal representation. This research turns the abstract problem of historical underrepresentation into specific engineering targets with clear data volume requirements for each decade. For example, our methodology can determine that data from the 1950s may be underrepresented by a factor of five compared to content from the 2020s, requiring a fivefold increase in historical training data for comparable performance.

This precise quantification enables evidence-based approaches to temporal data balancing through a mathematical framework that correlates token distributions with temporal representation. Our linear programming approach derives specific volume multipliers for each decade, indicating exactly how much additional data would be required to achieve balanced representation. These quantitative targets represent a significant advancement over current approaches, which often lack specific guidance for addressing representational imbalances.

1.2.3 Supporting Temporal Reasoning Applications

Language models are increasingly used in applications that need sophisticated temporal reasoning, including historical text analysis, forecasting, and temporal question answering. The effectiveness of these applications depends critically on how well models understand and represent different time periods. Temporal biases can lead to inconsistent performance across these applications, limiting their reliability and usefulness.

By providing insights into temporal distributions in training data, this research directly contributes to improving models for time-sensitive applications. Understanding current limitations in temporal representation is the first step toward developing more temporally robust models capable of consistent performance across different historical contexts.

1.2.4 Ensuring Scientific Validity in Research

The growing use of language models in scientific and academic research raises important questions about the validity of AI-assisted findings, especially when historical content is involved. Medical researchers analyzing historical case studies, clinical trials, or long-term health data using language models may draw incorrect conclusions if these models inadequately represent certain time periods. Similarly, humanities scholars studying historical texts through computational methods may encounter systematic biases in how these tools interpret period-specific language.

The reliability of AI-assisted scientific insights depends critically on how well models represent different historical periods. If models contain significant temporal biases, findings about historical periods may be fundamentally flawed without researchers being aware of these limitations. For example, a medical study using language models to analyze changing descriptions of symptoms over time might produce skewed results if the model's representation of medical terminology from earlier decades is inadequate.

This research provides crucial transparency about potential temporal biases that might affect scientific validity. By making these biases explicit and quantifiable, researchers can better assess the reliability of model-generated content about different time periods and take appropriate steps to address potential distortions or misrepresentations in their findings.

1.2.5 Advancing Model Auditing Methodologies

This research extends data mixture inference techniques to a new dimension, showing how tokenizer analysis can reveal previously hidden aspects of training data composition. By adapting linear programming approaches to identify decade-specific patterns in tokenizer merge rules, this project establishes a novel methodological framework for temporal analysis of language models.

The methodological innovation demonstrated in this work contributes to the broader field of AI auditing and transparency. The techniques developed allow researchers to infer important characteristics of training data without needing access to private datasets, advancing our collective ability to understand and evaluate commercial language models.

1.2.6 Enhancing AI Transparency and Accountability

Current documentation practices for language models rarely provide details about the temporal distribution of training data. This lack of transparency leaves users with insufficient information to assess model reliability for applications involving historical content. The problem is particularly acute for commercial models, where training data composition is often considered proprietary.

By developing techniques to infer temporal distribution without requiring access to the original training data, this research contributes to greater transparency and accountability in AI systems. These insights enable users to make more informed decisions about when and how to apply language models to historical texts, and allow for better assessment of whether specific models are suitable for temporally sensitive applications.

This enhanced transparency serves both practical engineering goals of improved model development and broader ethical objectives of ensuring that AI systems represent diverse historical perspectives rather than predominantly contemporary viewpoints.

1.3 Research Gap and Problem Statement

Despite growing research on language model biases, there's a significant gap in understanding the temporal distribution of training data and how this affects model performance. While much work has been done on demographic biases, gender representation, and multilingual capabilities, temporal aspects remain relatively unexplored. There is currently no established methodology for quantifying the temporal distribution of language model training data without direct access to the original corpus.

The problem is made worse by the lack of transparency in commercial language models, whose training methodologies and data sources are often proprietary. Users of these models have limited insight into potential temporal biases that might affect performance when processing texts from different historical periods. This lack of transparency makes it difficult to assess model reliability for applications involving historical content.

Furthermore, the relationship between temporal data distribution and model performance across different decades remains poorly understood. It is unclear whether models show performance disparities across time periods, and if so, whether these disparities correlate with the temporal distribution of their training data.

This research addresses these gaps by developing a methodology to infer the temporal distribution of language model training data through analysis of tokenizer patterns, and by investigating the relationship between this distribution and model performance across different historical periods.

1.4 Research Questions

This thesis investigates the following research questions:

RQ1: How is training data distributed across different decades? This question aims to quantify the proportion of training data from each decade by analyzing tokenizer merge patterns, adapting Hayase et al.'s linear programming methodology to temporal classification. This will reveal potential imbalances in historical representation within language models.

RQ2: What is the correlation between data volume and temporal recency? This question investigates whether more recent time periods are overrepresented in training data by examining the relationship between data quantity and decade of origin. This analysis will help identify potential recency bias in model training.

RQ3: How do temporal distributions affect model performance across different decades? This question explores how well models perform on texts from different time periods to understand the practical impact of temporal biases. This includes analyzing task-specific performance variations and error patterns across decades.

1.5 Thesis Objectives and Scope

The main goal of this thesis is to develop and validate a methodology for inferring the temporal distribution of language model training data through analysis of tokenizer patterns. Specifically, this research aims to:

First, adapt Hayase et al.ś data mixture inference approach to identify temporal patterns in tokenizer vocabularies. Second, create a comprehensive temporal reference dataset spanning from the 1850s to the 2020s. Third, quantify the temporal distribution of training data in several commercial language models. Fourth, investigate the relationship between temporal distribution and model performance across decades. Finally, provide insights and recommendations for addressing temporal biases in language model development.

The scope of this research is limited to English language content across different time periods. While multilingual temporal analysis would be valuable, focusing on English allows for more controlled experiments and more reliable historical data sources. The research focuses on BPE tokenizers, which are the predominant tokenization method in modern language models. The temporal range considered spans from the 1850s to the 2020s, divided into decade-long periods.

This research does not attempt to modify existing language models to address temporal biases, nor does it aim to create a temporally balanced language model. Rather, it focuses on developing methodologies for detecting and quantifying temporal biases in existing models.

1.6 Research Methodology Overview

This research uses a mixed-methods approach combining computational analysis with statistical validation. The methodology builds on Hayase et al.'s data mixture inference framework, adapting it to identify temporal patterns in tokenizer vocabularies.

The core approach involves analyzing the merge rules of BPE tokenizers, which reflect the frequency distribution of token pairs in the training data. By comparing these merge patterns against a temporal reference dataset with known decade distributions, we can infer the most likely temporal composition of the original training data.

The research methodology consists of several key components:

First, data collection and preprocessing involves creating a comprehensive temporal reference dataset spanning from the 1850s to the 2020s, using sources such as the British Library, Project Gutenberg, and more recent web content. Second, tokenizer analysis examines BPE merge rules to identify decade-specific patterns and distinctive vocabulary. Third, linear programming for distribution inference adapts Hayase et al.'s linear programming approach to solve for the temporal distribution that best explains

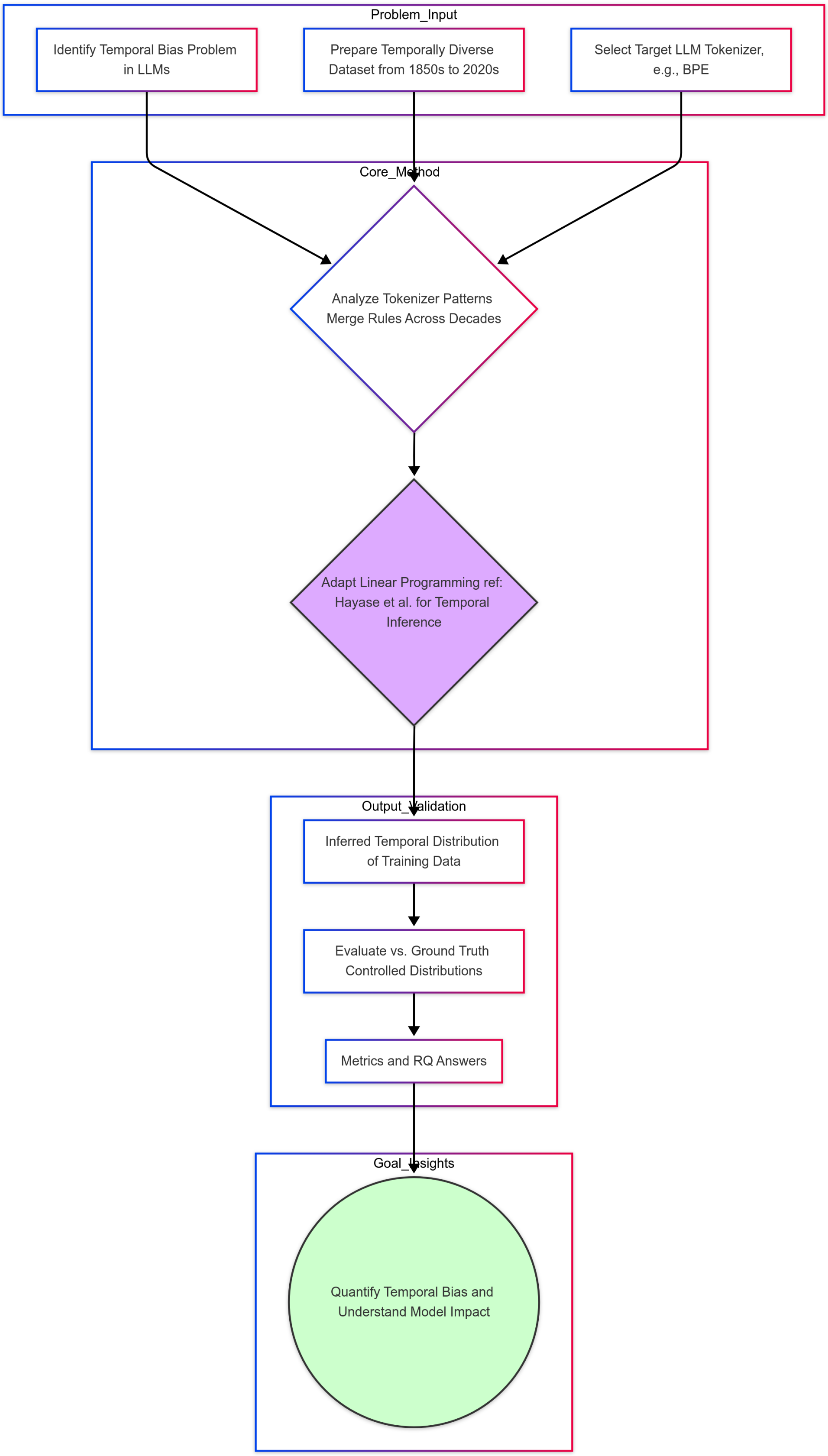

Problem_Input
Identify Temporal Bias Problem in LLMs
Prepare Temporally Diverse Dataset from 1850s to 2020s
Select Target LLM Tokenizer, e.g., BPE
Core_Method
Analyze Tokenizer Patterns Merge Rules Across Decades
Adapt Linear Programming ref: Hayase et al. for Temporal Inference
Output_Validation
Inferred Temporal Distribution of Training Data
Evaluate vs. Ground Truth Controlled Distributions
Metrics and RQ Answers
Goal_Insights
Quantify Temporal Bias and Understand Model Impact

Figure 1.1: High-level workflow of the temporal bias analysis methodology

observed tokenizer patterns. Fourth, statistical validation uses bootstrap analysis and other statistical methods to validate findings and quantify uncertainty. Finally, controlled experiments are conducted with known temporal distributions to validate the methodology before applying it to commercial language models.

This methodology enables us to infer temporal distributions without requiring direct access to the original training data, making it suitable for auditing commercial language models with proprietary training methodologies.

1.7 Thesis Contributions

This thesis makes several significant contributions to understanding temporal biases in language models:

First, it introduces a novel methodology for inferring the temporal distribution of language model training data through analysis of tokenizer patterns. This approach extends Hayase et al.'s data mixture inference framework from language distribution to temporal distribution, providing a valuable tool for auditing commercial language models.

Second, it creates a comprehensive temporal reference dataset spanning from the 1850s to the 2020s, which can serve as a benchmark for future research on temporal aspects of language models. This dataset combines historical texts from various sources with careful temporal labeling, addressing the challenge of creating reliable training data for different time periods.

Third, it provides the first quantitative analysis of temporal distribution in several commercial language models, revealing patterns of representation across different decades. These findings offer insights into potential temporal biases that may affect model performance in various applications.

Fourth, it investigates the relationship between temporal distribution and model performance, contributing to our understanding of how training data composition affects a model's ability to process texts from different time periods. This knowledge can inform more balanced approaches to language model development.

Finally, it proposes specific recommendations for addressing temporal biases in language model development, offering practical guidance for creating more temporally robust AI systems.

1.8 Defaults

This section provides a brief overview of default settings and assumptions used throughout the thesis, ensuring clarity and reproducibility.

Throughout this research, we use Byte-Pair Encoding (BPE) tokenizers with a vocabulary size of 50,265 tokens, following common practice in modern language model development. When not otherwise specified, we sample 1 GB of text per decade for calculating token pair frequencies, consistent with Hayase et al.'s methodology.

For statistical validation, we use bootstrap analysis with 50 iterations and a sample ratio of 0.8 by default. The default evaluation metric for measuring the accuracy of distribution inference is $\log_{10}$(Mean Squared Error), which allows direct comparison with Hayase et al.'s results.

The temporal range considered spans from the 1850s to the 2020s, divided into decade-long periods (1850s, 1860s, etc.). When analyzing commercial language models, we focus on their publicly available tokenizers rather than the model weights, which are often proprietary.

1.9 Options

This section outlines the various configurations and alternative approaches explored in this research, providing context for methodological choices.

While our primary methodology focuses on BPE merge rules, we also explored alternative approaches for inferring temporal distribution, including tokenizer encoding efficiency analysis and token classification methods. These alternatives serve as baselines for evaluating our primary approach.

For temporal reference datasets, we considered multiple sources, including the British Library, Project Gutenberg, the OSCAR corpus, and more recent web content. The final dataset combines these sources to achieve balanced representation across time periods, with different weighting strategies explored to address the inherent imbalance in historical data availability.

We investigated several data mixture scenarios in our controlled experiments, including uniform distribution, recency bias, historical bias, and bimodal distribution. These scenarios allow us to evaluate our methodology's performance under different temporal patterns.

In our analysis of commercial language models, we considered tokenizers from different model families, including GPT, LLAMA, MISTRAL and others. This diverse selection enables comparative analysis across model generations and development approaches.

For statistical validation, we explored multiple techniques, including bootstrap analysis, cross-validation and sensitivity analysis. These complementary approaches help establish the robustness of our findings and quantify uncertainty in our distribution estimates.

Chapter 2. Literature Review

2.1 Introduction to Language Models and Their Evolution

Large Language Models (LLMs) have evolved significantly since the introduction of the transformer by Vaswani et al. (2017). The transformer architecture, with its self-attention mechanisms, changed natural language processing by enabling parallel processing of input sequences while capturing long-range dependencies better than previous recurrent neural network approaches. This innovation has been refined in subsequent model generations, with each iteration introducing architectural improvements to enhance performance and efficiency.

Modern commercial LLMs like GPT-4 (OpenAI, 2023), LLAMA 3 (Touvron et al., 2024) and MISTRAL (Jiang et al., 2023) build on the transformer foundation while implementing proprietary modifications to attention mechanisms, normalization techniques and activation functions. These architectures typically scale to hundreds of billions of parameters organized across dozens or hundreds of layers, with each layer containing multiple attention heads that capture different linguistic patterns (Brown et al., 2020).

The scaling trajectory has been remarkable, with GPT-2's 1.5 billion parameters growing to GPT-3's 175 billion, leading to improvements in capabilities across various language tasks. While parameter scaling continues to be a dominant strategy, recent research has also focused on more efficient architectures. Notably, models like MISTRAL have shown that careful architectural choices can produce models that outperform much larger competitors, suggesting that raw parameter count is not the only path to better performance (Jiang et al., 2023).

Training large language models requires enormous text corpora. Commercial models analyzed by Longpre et al. (2023) typically train on hundreds of billions to trillions of tokens, drawn from diverse sources. Brown et al. (2020) documented the Common Crawl web data, WebText, books, and Wikipedia as primary components of GPT-3's training corpus, while more recent models include code repositories, academic papers, and specialized domain-specific datasets (Touvron et al., 2023).

The quality and composition of these datasets significantly influence model behavior. Research by Dodge et al. (2021) and Longpre et al. (2023) shows that data quality often matters more than quantity, with models performing better when trained on carefully filtered corpora. Similarly, domain coverage directly impacts a model's capabilities in specific areas–models trained with substantial code components, for instance, show enhanced reasoning abilities (Wei et al., 2022).

While much attention has been paid to linguistic diversity and domain composition of training data, the temporal dimension has received less focus. The distribution of texts across different historical periods potentially affects a model's ability to understand language from different eras, yet this aspect remains understudied despite its importance for applications in historical text analysis and cultural heritage preservation.

2.2 Temporal Aspects in Language Processing

2.2.1 Linguistic Drift and Language Evolution

Language naturally evolves over time, with new terms emerging, existing words taking on new meanings, and grammatical conventions shifting (Dubossarsky et al., 2017; Hamilton et al., 2016; Xu and Kemp, 2018). This linguistic drift poses significant challenges for natural language processing systems trained mainly on contemporary data. The phenomenon is well-documented in linguistics research but has only recently begun to receive attention in computational linguistics (Del Tredici et al., 2019).

Research on linguistic drift typically examines changes in word meaning and usage patterns over time. For instance, terms like "web," "cloud," and "tweet" have taken on fundamentally different primary meanings in the digital era compared to their historical usage (Hamilton et al., 2016). These semantic shifts, along with syntactic and stylistic evolution, create substantial differences between historical and contemporary texts (Dubossarsky et al., 2017).

The temporal dimension of language introduces a unique challenge for language models. Unlike static domains, language continues to evolve even after a model is trained, potentially leading to a form of temporal drift where model performance gradually degrades for processing contemporary language. Conversely, models trained mainly on contemporary data may struggle to correctly interpret historical texts, showing the bidirectional challenge of temporal generalization.

2.2.2 Challenges in Processing Historical Texts

Processing historical texts introduces several distinct challenges for NLP systems. Historical language patterns may be tokenized less efficiently than contemporary ones, leading to performance disparities. This inefficiency stems from both vocabulary changes and stylistic evolution (Baron and Rayson, 2008; Schneider and Hovy, 2018).

Similarly, syntactic patterns in historical texts often feature more complex clause structures, different word order conventions, and formulations that would be considered archaic by contemporary standards (Baron and Rayson, 2008; Schneider and Hovy, 2018). Historical texts often use different grammatical conventions, vocabulary, and orthographic patterns compared to modern writing (Baron and Rayson, 2008). For instance, English texts from the 19th century frequently use longer sentences, more complex clause structures, and vocabulary terms that have since fallen out of common usage (Schneider and Hovy, 2018). These differences can challenge models trained primarily on contemporary data, affecting tasks from basic parsing to higher-level understanding.

The technological and cultural references in historical texts also differ substantially from contemporary ones. Models trained on modern data may struggle to correctly contextualize references to historical technologies, social practices, or cultural phenomena that are rare or nonexistent in contemporary discourse. These challenges highlight the importance of temporal diversity in training data for models expected to process texts from different historical periods.

2.2.3 Impact on Model Performance

Temporal biases in language models can significantly affect their performance across different historical contexts. When models are trained mainly on contemporary data, they may develop a form of "presentism" in their language understanding capabilities. This bias shows up as varying performance levels when processing texts from different time periods.

Several studies have observed performance disparities across temporal contexts. Lazaridou et al. (2021) documented how language models struggle with temporal adaptation, showing degraded performance on tasks involving rapidly evolving domains. More specifically, Bjerva et al. (2019) showed that neural language models perform significantly better on modern texts than historical ones, even when controlling for domain and complexity. These performance gaps suggest an underlying temporal bias, potentially stemming from imbalanced representation in training data.

The performance impact extends beyond simple comprehension to more nuanced aspects of language understanding. Models trained mainly on contemporary data may struggle to correctly interpret historical idioms, references, and contextual implications. Conversely, they may inappropriately apply contemporary connotations to historical terms, leading to misinterpretations of historical texts.

As language models are increasingly used for analyzing historical texts in digital humanities projects, these temporal performance disparities become particularly concerning. Applications in historical text analysis, cultural heritage preservation, and diachronic linguistics all depend on reliable processing of texts from different time periods. Understanding and addressing temporal biases is therefore essential for ensuring the reliability of these applications.

2.2.4 Ethical Considerations

Temporal biases in AI systems raise important ethical considerations regarding the representation of different historical periods and perspectives (Crawford, 2021; Metzler et al., 2021). If language models disproportionately represent contemporary language patterns, they may inadvertently marginalize historical viewpoints and perpetuate temporal inequity in AI systems (Metzler et al., 2021).

This bias can manifest in several ways:

- First, it may lead to inaccurate representations of historical contexts and viewpoints in model-generated content (Crawford, 2021). When prompted to produce text about historical periods, models biased toward contemporary language patterns may impose present-day sensibilities, vocabulary and framing on historical topics, resulting in anachronistic content.

- Second, temporal biases may affect the preservation and interpretation of cultural heritage (Metzler et al., 2021). As historical texts are increasingly digitized and processed by AI systems, models with temporal biases may introduce systematic errors in their interpretation, potentially distorting our understanding of historical documents and cultural artifacts.

- Finally, there are broader implications for epistemic justice, the fair distribution of epistemic goods

> such as knowledge and understanding (Crawford, 2021). If AI systems consistently perform better on contemporary texts than historical ones, this creates an imbalance in the accessibility and processing of knowledge from different time periods, potentially reinforcing existing temporal hierarchies in knowledge production and dissemination.

Addressing temporal biases in language models is therefore not merely a technical challenge but an ethical imperative for ensuring fair representation of different historical periods and perspectives in AI systems.

2.3 Data Distribution Analysis Techniques

2.3.1 Tokenization Fundamentals

Tokenization is a fundamental preprocessing step in language model training, converting raw text into sequences of tokens from a fixed vocabulary. Byte-Pair Encoding (BPE), first applied to NLP by Sennrich et al. (2016) and now the dominant approach, uses a data-driven algorithm to iteratively merge the most frequent adjacent byte pairs in the training corpus.

As described in Hayase et al. (2024), the BPE algorithm starts with a vocabulary of individual bytes or characters and iteratively identifies the most frequent adjacent pairs, adding them to the vocabulary as new tokens. This process continues until the desired vocabulary size is reached, resulting in an ordered list of merge rules. These merge rules are then applied in sequence to tokenize new text, breaking down words into subword units.

This data-driven approach to vocabulary construction means that the resulting tokenizer reflects the statistical patterns of the training data. As Hayase et al. (2024) show, this property makes BPE tokenizers particularly valuable for inferring properties of the training data, including its distributional makeup.

The choice of tokenization strategy significantly impacts model performance, particularly across different data domains and languages. As explored by Ács (2019) in "Exploring BERT's Vocabulary," tokenization efficiency varies dramatically across languages, with some languages experiencing much higher "fertility" (the ratio of tokens to words) than others. This disparity can lead to uneven representation in multilingual models and varying performance across languages.

Recent work by Ahia et al. (2023) shows that tokenization disparities can lead to significant cost differences when using commercial language models across languages, with some languages requiring substantially more tokens per sentence than others. This tokenization inefficiency creates both economic and performance disparities across languages.

The relationship between tokenizer design and model performance extends to temporal contexts as well. Historical language patterns may be tokenized less efficiently if they are underrepresented in the training data, potentially leading to degraded performance on historical texts. However, this temporal aspect of tokenization has been less thoroughly investigated than cross-lingual effects.

2.3.2 Data Mixture Inference Approaches

The methodological foundation for this thesis builds directly on the groundbreaking work by Hayase et al. (2024) titled "Data Mixture Inference: What do BPE Tokenizers Reveal about their Training Data?" This research introduced a novel approach to inferring the distributional makeup of training data through analysis of BPE tokenizer merge rules, providing a significant contribution to language model auditing without requiring access to the original training data or model weights.

The key insight of Hayase et al.'s work is that the ordered list of merge rules learned by a BPE tokenizer reveals information about token frequencies in the training data. During tokenizer training, the BPE algorithm iteratively finds the most frequent pair of tokens in the corpus, adds it to the merge list, and applies it to the dataset. This process creates an ordered list of merge rules that inherently contains information about the statistical patterns in the training data.

The authors formulate this insight as a linear programming problem where the objective is to find the mixture of reference datasets that would produce merge patterns most similar to those observed in the target tokenizer. By solving this optimization problem, they infer the proportions of different languages, programming languages, and data sources in the tokenizer's training data.

In controlled experiments, Hayase et al. demonstrated high precision, with $\log_{10}$(MSE) values of -7.30±1.31 when recovering known mixture ratios. Applying this approach to commercial tokenizers, they uncovered previously unknown information about model training data, including GPT-4O being trained on 39% non-English text, MISTRAL NEMO on 47% non-English text, and LLAMA 3 on 48% non-English text. These findings demonstrate both the technical validity and practical utility of tokenizer-based inference for auditing language model training data.

While Hayase et al. focused on language and domain distribution, the same foundational insight, that tokenizer merge rules reflect statistical patterns in training data, can be extended to temporal distribution. This thesis adapts their linear programming approach to solve for temporal mixtures rather than language mixtures, applying similar mathematical principles to a different dimension of data analysis.

2.3.3 Other Model Auditing Techniques

Beyond Hayase et al.'s tokenizer-based approach, several other methods have been developed for auditing language models and inferring properties of their training data. Membership inference attacks, as explored by Carlini et al. (2021), Carlini et al. (2022), and Shokri et al. (2017), attempt to determine whether specific examples were part of a model's training data. While these techniques focus on individual instances rather than broader distributions, they share the goal of inferring properties of training data without direct access.

Distribution inference attacks, which aim to uncover global properties of training data, have been studied across various machine learning contexts. Early works by Ateniese et al. (2015) demonstrated attacks against machine learning classifiers, while later research extended these approaches to convolutional neural networks (Ganju et al., 2018) and GANs (Zhou et al., 2022). These approaches typically use a meta-classifier strategy, training models on datasets with different properties and then using the resulting behaviors to infer the properties of target models.

For language models specifically, recent work by Duan et al. (2024) evaluated the effectiveness of various membership inference attacks, finding that many techniques perform near random when pretraining data is deduplicated. This highlights the challenge of developing effective auditing techniques for modern language models, which often use advanced data cleaning and deduplication procedures.

The temporal distribution inference approach developed in this thesis differs from previous methods in that it leverages the internal structure of BPE tokenizers rather than model outputs or behaviors. This makes it particularly suitable for analyzing commercial language models where only the tokenizer (not the model weights) may be publicly available. Additionally, by focusing on temporal distribution rather than membership or demographic properties, this work addresses an understudied dimension of model auditing with significant implications for understanding model performance across different historical contexts.

2.4 Research Gaps and Current Work

Despite the growing body of research on language model biases, there exists a significant gap in understanding the temporal distribution of training data and its effects on model performance. While extensive work has examined demographic biases (Blodgett et al., 2020), gender representation (Bolukbasi et al., 2016; Zhao et al., 2018) and multilingual capabilities (Joshi et al., 2020), temporal aspects remain relatively unexplored.

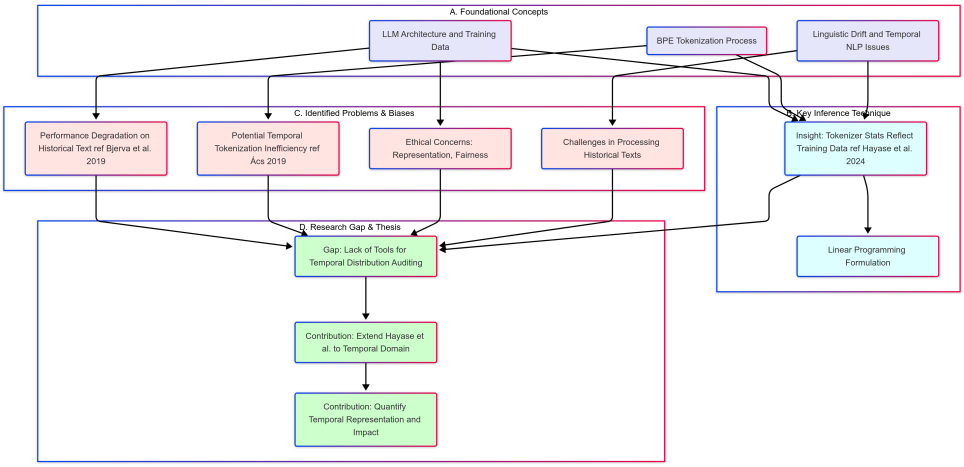

A. Foundational Concepts
LLM Architecture and Training Data
BPE Tokenization Process
Linguistic Drift and Temporal NLP Issues
C. Identified Problems & Biases
Performance Degradation on Historical Text ref Bjerva et al. 2019
Potential Temporal Tokenization Inefficiency ref Ács 2019
Ethical Concerns: Representation, Fairness
Challenges in Processing Historical Texts
B. Key Inference Technique
Insight: Tokenizer Stats Reflect Training Data ref Hayase et al. 2024
Linear Programming Formulation
D. Research Gap & Thesis
Gap: Lack of Tools for Temporal Distribution Auditing
Contribution: Extend Hayase et al. to Temporal Domain
Contribution: Quantify Temporal Representation and Impact

Figure 2.1: Conceptual framework showing the relationship between different aspects of language model research and the positioning of temporal bias analysis

The lack of research on temporal biases is particularly notable given the increasing application of language models to historical text analysis in fields such as digital humanities and computational social science. As these models are deployed to process and analyze historical texts, understanding their temporal capabilities and limitations becomes crucial for ensuring reliable results.

Several factors contribute to this research gap:

- First, there is a lack of standardized datasets for evaluating temporal capabilities. Unlike domains such as computer vision or machine translation, which have established benchmarks spanning different time periods, NLP has few resources specifically designed to assess performance across historical contexts. This makes it difficult to systematically evaluate and compare models along the temporal dimension.

- Second, the proprietary nature of many commercial language models limits access to information about their training data composition, including temporal distribution. While some model developers provide high-level descriptions of their training data sources, detailed information about the distribution across time periods is rarely disclosed, making it challenging to analyze potential temporal biases without specialized inference techniques.

- Finally, the methodological challenge of inferring temporal properties without direct access to training data has impeded progress in this area. Until recently, there were few effective techniques for auditing language models to determine the temporal composition of their training data, limiting researchers' ability to investigate this dimension of model bias.

This thesis addresses these gaps by developing a novel methodology for inferring the temporal distribution of language model training data through analysis of tokenizer patterns. By adapting Hayase et al.'s data mixture inference approach from language distribution to temporal distribution, this research provides a tool for auditing commercial language models without requiring access to their training data.

Furthermore, this work creates a comprehensive temporal reference dataset spanning from the 1850s to the 2020s, addressing the lack of standardized resources for temporal analysis. By investigating the relationship between temporal distribution and model performance, this research contributes to our understanding of how training data composition affects a model's ability to process texts from different time periods.

In positioning this work within the broader research landscape, this thesis bridges the gap between language model auditing techniques and temporal analysis in natural language processing. It extends existing approaches for inferring properties of training data to the temporal dimension, offering a new perspective on language model biases and capabilities. The methodology and findings presented here have implications for model development, evaluation and application, particularly in contexts involving historical text analysis and cultural heritage preservation.

Chapter 3. Methodology

This chapter describes the methods used to study temporal biases in Large Language Model (LLM) training data. Building on Hayase et al. (2024)'s work on data mixture inference using BPE tokenizers for language identification, we adapt these techniques to study temporal distribution. We test the adapted BPE/LP method using controlled experiments with synthetic datasets that have known temporal distributions (uniform, recency biased, historically biased and bimodal). The results show that while the method can detect different temporal distributions, its accuracy in recovering these distributions (with $\log_{10}$(Mean Squared Error) values between -1.8 and -2.2) is much lower than reported for language inference, showing the difficulty of capturing gradual language changes through merge rule analysis.

The method extends the BPE/LP inference framework from language identification to temporal distribution analysis, addressing the challenges of detecting gradual language change over time. It introduces new techniques for handling temporal noise and common token interference in merge rule analysis. The analysis consistently finds patterns of over-representation for older historical periods and under-representation for recent decades across different models and tokenizer types, providing insights into the limitations of this inference approach for temporal analysis and highlighting potential biases affecting model performance.

3.1 Methodological Challenges and Approach

Unlike language identification, where different vocabularies and grammar create clear statistical differences in token patterns, temporal analysis within a single language presents significant challenges. This thesis studies these challenges directly, finding that while BPE/LP methods can identify broad temporal patterns, they achieve much lower accuracy ($\log_{10}$(MSE) values between -1.8 and -2.2) than the language inference benchmark (-7.3).

Rather than claiming precise quantification of temporal distributions, this research focuses on characterizing the systematic bias patterns that emerge when applying tokenizer analysis to temporal data. The consistent over-representation of historical periods and under-representation of recent decades across different models and experimental conditions represents a key finding that advances our understanding of the limitations and potential of tokenizer-based inference methods.

3.2 Research Framework Overview

Inferring the temporal composition of opaque LLM training data presents unique difficulties compared to inferring mixtures of different languages. Language change across decades is often gradual and subtle, lacking the clear vocabulary and structural boundaries that separate different languages. This research studies whether the statistical patterns in BPE tokenizer merge rules, reflecting frequent token co-occurrences during training, can serve as a proxy for the temporal distribution of the underlying training corpus.

The methodology, shown in Figure 3.1, involves input preparation, data preprocessing, tokenizer analysis, temporal distribution inference, and evaluation. The process starts with selecting a target LLM tokenizer and preparing temporally diverse reference text corpora spanning from the 1850s to 2020s. For controlled experiments, we define target ground truth temporal distributions such as Uniform, Recency Bias, Historical Bias, and Bimodal distributions.

The data preparation phase involves creating controlled datasets for tokenizer training and sampling reference corpora for frequency analysis, ensuring temporal labeling and balanced representation where possible. The tokenizer analysis stage examines the target tokenizer's merge rules against decade-specific reference corpora to calculate token pair frequencies at each merge step, using noise reduction techniques like token removal.

The temporal distribution inference uses Linear Programming (LP) to estimate the decade proportions (α_i) that best explain the observed merge rule sequence based on the calculated frequencies. Finally, the evaluation phase compares the inferred distribution against known ground truths in controlled experiments using various metrics, identifying systematic biases and assessing robustness.

3.3 Experimental Setup

Rigorous evaluation requires controlled experiments where the ground truth is known. This allows for quantitative assessment of the methodology's accuracy and limitations before applying it to commercial tokenizers with unknown training data composition.

3.3.1 Controlled Experiments Design

Four distinct data mixture scenarios, representing plausible or informative temporal distributions, were designed for evaluating the inference method. The Uniform Distribution serves as an idealized baseline where each decade (1850s to 2020s) contributes equally ($\sim$5.9%), testing the recovery of a perfectly balanced temporal signal. The Recency Bias scenario mimics assumed modern LLM training data, with proportions weighted heavily towards recent decades (1990s to 2010s). The Historical Bias scenario represents an inverted case emphasizing older content (1850s to 1960s), testing performance on historically skewed data. Finally, the Bimodal Distribution features peaks in both early (1850s) and recent (2010s to 2020s) periods, testing the ability to detect multiple temporal concentrations.

Synthetic datasets were constructed by sampling from reference corpora according to these exact proportions (details in Section 3.4). Tokenizers (Vocabulary Size: 32,768, standard BPE) were trained on these datasets under consistent hyperparameters. Our experiments showed that while the method successfully detects broad temporal patterns, its quantitative accuracy ($\log_{10}$(MSE) between -1.8 and -2.2) indicates significant challenges in precise distribution recovery, particularly compared to the original language inference task.

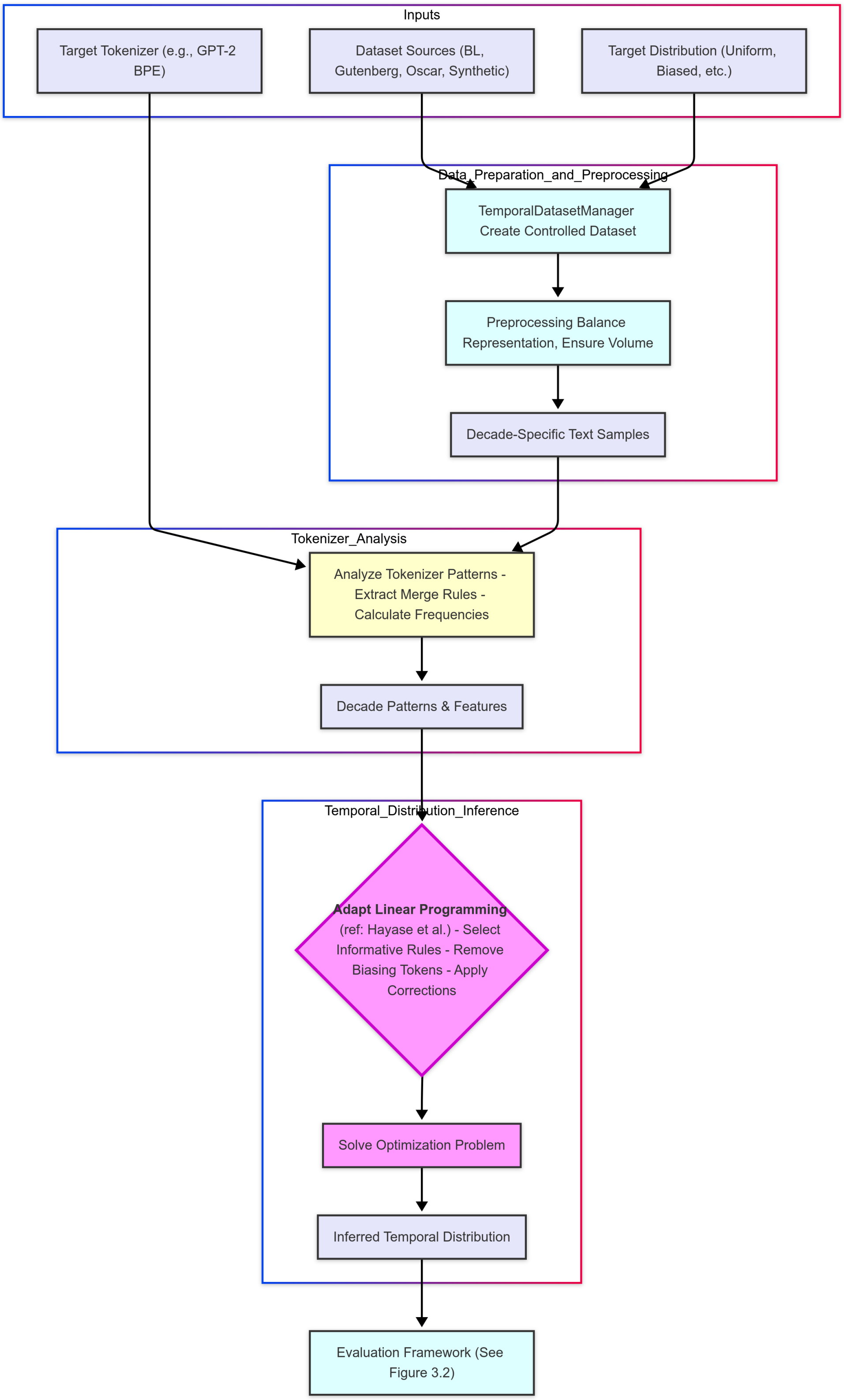

Inputs
Target Tokenizer (e.g., GPT-2 BPE)
Dataset Sources (BL, Gutenberg, Oscar, Synthetic)
Target Distribution (Uniform, Biased, etc.)
Data_Preparation_and_Preprocessing
TemporalDatasetManager Create Controlled Dataset
Preprocessing Balance Representation, Ensure Volume
Decade-Specific Text Samples
Tokenizer_Analysis
Analyze Tokenizer Patterns - Extract Merge Rules - Calculate Frequencies
Decade Patterns & Features
Temporal_Distribution_Inference
Adapt Linear Programming (ref: Hayase et al.) - Select Informative Rules - Remove Biasing Tokens - Apply Corrections
Solve Optimization Problem
Inferred Temporal Distribution
Evaluation Framework (See Figure 3.2)

Figure 3.1: Workflow of the temporal bias analysis methodology

3.3.2 Evaluation Metrics

Method performance was evaluated by comparing the inferred distribution ($\hat{\alpha}$) against the known ground truth distribution (α^*) using several metrics (summarized in Table 3.1). The primary metric, Log_{10}(Mean Squared Error) (Log_{10}(MSE)), calculated as $\log_{10}(\text{mean}((\hat{\alpha}_i - \alpha_i^*)^2))$, follows Hayase et al. (2024) and allows direct comparison. Lower values indicate better accuracy, with the $\log_{10}$ scale facilitating comparison of errors potentially spanning orders of magnitude. Additional metrics include Mean Absolute Error (MAE), measuring the average magnitude of error per decade; Jensen Shannon Distance (JSD), a bounded [0, 1] information theoretic measure of similarity between distributions; and Spearman's Rank Correlation, measuring if the relative ranking of decades by proportion is preserved.

Metric	Description	Range
Log_{10}(MSE)	Logarithm of Mean Squared Error	$(-\infty, 0]$
MAE	Mean Absolute Error	[0, 1]
JSD	Jensen Shannon Distance	[0, 1]
Spearman's ρ	Rank Correlation	[-1, 1]

Table 3.1: Summary of Primary Evaluation Metrics for Temporal Distribution Inference

3.4 Data Collection and Preprocessing

A comprehensive temporal reference corpus spanning 1850s to 2020s was essential for this research. The corpus was compiled from diverse sources including the British Library's digitized archives, Project Gutenberg, and the OSCAR web corpus. This corpus serves multiple purposes: providing ground truth data for controlled experiments, enabling frequency analysis for LP inference, and supporting validation of the methodology's assumptions about temporal linguistic patterns.

The source integration process combined texts from various origins to maximize coverage across decades. The British Library's digitized archives provided historical texts from the 1850s-1950s, while Project Gutenberg contributed literary works across the time period. The OSCAR web corpus supplied recent content from the 1990s-2020s, supplemented by additional specialized collections for specific decades or genres.

Temporal attribution involved rigorous verification of publication dates using metadata from source collections, computational historical linguistic techniques, cross-validation with multiple sources, and expert review for ambiguous cases. The preprocessing phase included standard text cleaning and normalization: UTF-8 normalization, removal of non-textual elements, standardization of formatting, and handling of OCR errors in historical texts.

For frequency analysis, representative samples were drawn from each decade's collected texts, targeting a volume of 1GB to 1.5GB per decade where possible. The sampling process maintained genre balance where feasible, used stratified sampling to ensure representation of different text types, and included quality checks to verify temporal consistency. Achieving balanced representation across all decades proved challenging, particularly for earlier periods where digitized materials are less abundant. This data sparsity issue was addressed through careful sampling strategies and quality control measures, though it remains a limitation acknowledged in the Discussion chapter.

3.5 Tokenizer Analysis and Inference

This section details the core BPE/LP adaptation for temporal inference.

3.5.1 BPE Merge Rules as Temporal Signal

The foundation of the inference method lies in the BPE algorithm's mechanism. By iteratively merging the most frequent adjacent token pair, the resulting ordered merge list encodes statistical information about the training data. While Hayase et al. (2024) used this for language identification, we hypothesize that subtle shifts in language use over time (vocabulary, common phrases, perhaps even style) will alter token pair frequencies, thus influencing the merge order and potentially leaving a recoverable temporal signature.

This approach does not rely on pre-defining specific lexical items as markers for particular decades. Instead, it uses the emergent statistical patterns captured by the entire sequence of BPE merge rules, assuming these aggregate patterns reflect the underlying temporal mixture of the training data. The challenge, distinct from language inference, lies in whether these structural merge patterns provide a sufficiently strong and unambiguous signal for subtle, continuous temporal change within a single language.

Isolating purely temporal linguistic features within historical corpora is inherently challenging, as changes in topic, genre, and style often co-occur with time, potentially confounding the frequency signals measured. Our methodology addresses this challenge through careful statistical analysis and validation, while acknowledging these inherent limitations in the Discussion chapter.

3.5.2 Token Frequency Calculation

To perform inference on a target tokenizer (e.g., GPT-2), we need to estimate how frequently its merge rules would occur in text from each specific decade. This involves extracting the ordered merge list $M = (m^{(1)}, m^{(2)}, ..., m^{(T)})$ from the target tokenizer. For each decade i in our reference corpus, we simulate the BPE process by applying merges $m^{(1)}$ through $m^{(t-1)}$ to the decade's text. At each step t, we calculate the frequency $c_{i,p}^{(t-1)}$ of all potential pairs p within the partially merged text of decade i. This computationally intensive step provides the necessary frequency data $c_{i,p}^{(t)}$ for the LP formulation.

3.5.3 Noise Reduction: Token Removal

Preliminary analysis of BPE merge rule frequencies across decades revealed a significant challenge: the frequency calculations and subsequent LP inference were often dominated by a small number of extremely frequent tokens. Manual inspection confirmed that many of these top tokens were either non-linguistic artifacts (such as HTML-like tags likely introduced during web corpus processing) or ubiquitous linguistic units (very common character pairs or function words/affixes that occur with high frequency across nearly all decades and genres, providing little discriminative temporal information).

These highly frequent, often temporally non-specific tokens act as statistical noise. Their overwhelming frequency can obscure the subtler frequency variations in less common, but potentially more temporally informative, merge pairs. This noise can bias the LP solution towards patterns driven by these common artifacts rather than genuine linguistic shifts over time.

To mitigate this noise and enhance the signal quality for the LP inference, a constrained token removal strategy was implemented as a preprocessing step after frequency calculation but before LP optimization. The process involved identifying biasing tokens, protecting distinctive tokens, and implementing constrained removal. For each analysis run, tokens were ranked based on a metric designed to capture their potential to introduce noise or bias. The top N=40 tokens were identified as candidates for removal, while a separate analysis identified 76 tokens as ’decade-distinctive’ that were protected from removal, regardless of their overall frequency.

This constrained filtering procedure proved essential for obtaining meaningful results. The removal of high-frequency noise, particularly non-linguistic artifacts like HTML tags, while preserving tokens potentially carrying valuable temporal signals, led to quantifiable improvements in accuracy metrics compared to initial runs performed without any token filtering. While not eliminating the fundamental challenges of temporal inference with BPE/LP, this step was necessary for obtaining meaningful insights from the available signals.

3.5.4 Linear Programming Formulation

Building on Hayase et al. (2024)’s approach for language mixture inference, we formulate the temporal distribution inference task as a specialized Linear Programming (LP) problem. The goal is to determine the set of decade proportions, $\alpha = (\alpha_{1850s}, \ldots, \alpha_{2020s})$, that best explains the observed merge sequence M (up to merge T, typically 30,000) of a target tokenizer, given the decade-specific frequency estimates $c_{i,p}^{(t-1)}$ derived from reference corpora.

The LP formulation includes several key components. The primary variables α_i represent the unknown, non-negative proportion of the training data originating from decade i. Core constraints ensure the proportions constitute a valid probability distribution ($\sum_i \alpha_i = 1$), maintain non-negativity ($\alpha_i \geq 0$ for all decades i), and enforce frequency dominance for each observed merge.

Recognizing that temporal distributions differ significantly from language mixtures, several novel constraints were added based on linguistic principles and empirical observations. These include minimum decade proportion constraints to prevent decades from being entirely excluded, maximum decade proportion constraints to counteract observed biases, and temporal smoothness constraints to reflect the gradual nature of language change.

The objective function seeks to maximize a weighted data-fit term, optionally augmented with a temporal trend regularizer. This objective aims to find the α that best aligns with the frequency evidence, particularly from merge rules deemed more informative or temporally relevant. The implementation uses the CVXPY optimization library with multiple robust solvers, including ECOS, SCS, and OSQP, with appropriate parameters to ensure convergence.

3.6 Implementation and Baselines

The core pipeline was implemented in Python using libraries like Transformers/Tokenizers (for BPE processing), NumPy/Pandas (for data handling) and CVXPY with the Gurobi solver for the LP optimization. Lazy constraint generation was used to manage the large scale of the LP problem.
For comparative context, baseline methods were implemented including Tokenizer Encoding Efficiency (TEE) based on byte to token ratios, Token Classification (TC) based on assigning vocabulary tokens to their most frequent decade, and Random Distribution using uniform sampling from the unit simplex.

Chapter 4. Experimental Setup

This chapter describes the experimental framework, data scenarios, processing pipeline and validation strategy used to evaluate the adapted BPE/LP method for temporal distribution inference, as described in Chapter 3. The setup is designed to test the method's capabilities and limitations under controlled conditions before analyzing commercial language model tokenizers.

4.1 Experimental Framework

The core experimental approach involved training BPE tokenizers on synthetic datasets with known temporal distributions and then applying our inference pipeline to assess how accurately the original distribution could be recovered. Key components of the experimental infrastructure and configuration are summarized in Table 4.1.

Component	**Configuration**
Infrastructure	Maxwell HPC: 8 cores, 64GB RAM, 100GB local storage
Software Stack	Python 3.9.12, Transformers (4.28.1), Tokenizers (0.13.2), CVXPY (1.3.1), Gurobi (10.0.0), NumPy, Pandas, Matplotlib, Seaborn
Dataset Size	Tokenizer Training: approximately 10 to 12GB total per scenario Frequency Estimation: 1GB to 1.5GB per decade
Tokenization	HuggingFace BPE implementation Vocabulary Size: 32,768 tokens Minimal preprocessing 100% Basic Multilingual Plane coverage
LP Configuration	Merges Considered (T): 30,000 Constraint Tolerance: 1e-6 Optimality Tolerance: 1e-5 Regularization (λ): 0.2 Lazy Constraint Generation
Validation	Bootstrap Analysis: 50 iterations 80% sampling ratio 95% Confidence Intervals

Table 4.1: Experimental Infrastructure Overview

4.1.1 Bootstrap Analysis Configuration

The bootstrap analysis used 50 iterations as a balance between providing an initial estimate of stability/uncertainty and managing computational resources for this exploratory study. This configuration was chosen based on the following considerations:

- **Resource Constraints:** Each iteration requires significant computational resources for frequency calculation and LP optimization
- **Statistical Power:** 50 iterations provide sufficient samples to estimate confidence intervals while maintaining reasonable runtime
- **Exploratory Nature:** As an initial investigation, this number of iterations provides preliminary insights into the method's stability

For future work aiming to generate publication-level confidence intervals, a larger number of iterations (e.g., 1000+) would be recommended to achieve more precise estimates of the uncertainty in the inferred distributions.

4.2 Distribution Analysis Implementation

To systematically evaluate the inference method, we implemented distinct data mixture scenarios and a standardized processing pipeline.

4.2.1 Data Mixture Scenarios

Four temporal distribution scenarios were designed to test the methodology under different conditions, mirroring the descriptions in Section 3.3.1:

First, the Uniform Distribution scenario ensures each decade from the 1850s to 2020s contributes an equal proportion (approximately 5.9%) of the data. This serves as a baseline for balanced temporal representation.

Second, the Recency Bias scenario features proportions that increase significantly in recent decades (peaking 1990s to 2010s), mimicking patterns often observed or assumed in modern LLM training data.

Third, the Historical Bias scenario weights proportions towards earlier decades (peaking 1850s to 1960s), testing the method's ability to handle data skewed towards historical content.

Fourth, the Bimodal Distribution scenario features distinct peaks in both historical (e.g., 1850s) and recent (e.g., 2010s to 2020s) periods, simulating a mixed training corpus and testing the ability to detect multiple temporal concentrations.

For each scenario, synthetic datasets were constructed by sampling from our comprehensive temporal reference corpus (detailed in Section 3.4) according to these precise proportions at the byte level, ensuring accurate ground truth representation.

4.2.2 Processing Pipeline

The core analysis followed a consistent pipeline (summarized in Table 4.2):

This pipeline ensures that the BPE/LP inference method is applied consistently across all controlled experiments.

Stage	Process	Key Aspects / Optimization
1. Tokenizer Training	Train scenario specific BPE tokenizers	Consistent parameters, vocab size
2. Merge Rule Extraction	Extract ordered merge rules	Handle different tokenizer formats
3. Frequency Analysis	Calculate decade specific frequencies	Parallel processing per decade
4. LP Inference	Apply Linear Programming model	Lazy constraint generation
5. Noise Reduction	Apply token removal strategy	Filter uninformative tokens
6. Evaluation	Compare inferred to ground truth	MSE, MAE, JSD metrics

Table 4.2: Processing Pipeline Architecture

4.3 Commercial Model Analysis

Following validation in controlled settings, the inference methodology was applied to analyze the publicly available tokenizers of several widely used commercial and open source language models.

4.3.1 Model Coverage

The analysis included tokenizers associated with the following models, covering different architectures and primary developers (as listed in Table 4.3):

Family	Models Analyzed	Primary Tokenizer Type(s)
GPT	GPT 2, GPT 2 Medium	BPE
BERT	BERT Base Uncased, ALBERT Base v2, ELECTRA Small Discriminator, DistilBERT Base Uncased	WordPiece
RoBERTa	RoBERTa Base, XLM RoBERTa Base	BPE, SentencePiece
Other	T5 Small, DistilGPT 2	SentencePiece, BPE

Table 4.3: Commercial Model Analysis Coverage

This selection provides a representative sample across common model families and tokenizer technologies (BPE, SentencePiece, WordPiece).

4.3.2 Validation Framework

Since ground truth is unknown for commercial models, validation involved: 1) applying the validated BPE/LP method, 2) identifying patterns in inferred distributions, 3) comparing results across tokenizer types, and 4) qualitatively assessing against external knowledge where possible.
The primary goal for commercial model analysis was identifying robust temporal representation patterns revealed by the inference method, rather than absolute accuracy measurement.

Chapter 5. Results and Analysis

This chapter presents the results from applying our adapted temporal distribution inference method, which uses BPE merge rule analysis and linear programming (LP), to various language models under controlled conditions. We evaluate the method's ability to recover known temporal distributions and identify potential biases in commercial tokenizers.

5.1 Overall Performance and Methodological Limitations

We tested the BPE/LP inference pipeline on four ground truth scenarios: Uniform, Recency Bias, Historical Bias and Bimodal. We analyzed ten different language models, covering various architectures and tokenizer technologies (BPE, SentencePiece, WordPiece).

Figure 5.1 shows how our method performed across different models and distribution types. The figure compares the inferred distributions (colored lines) with the known ground truth (black line) for each test scenario. The bottom row shows the absolute error (|Inferred minus Ground Truth|) per decade for each model, highlighting the discrepancies between inferred and actual distributions.

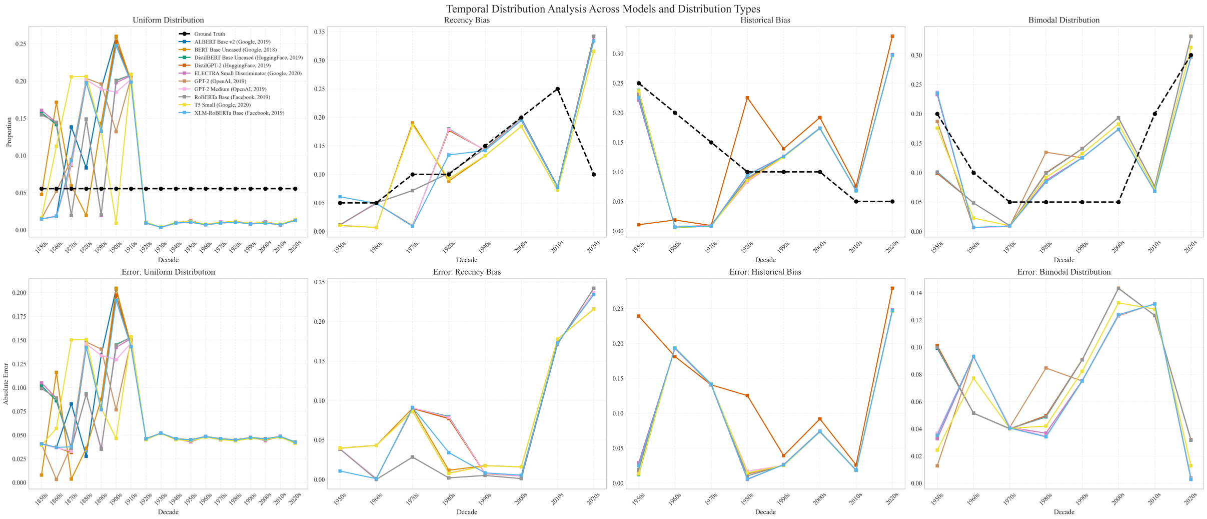
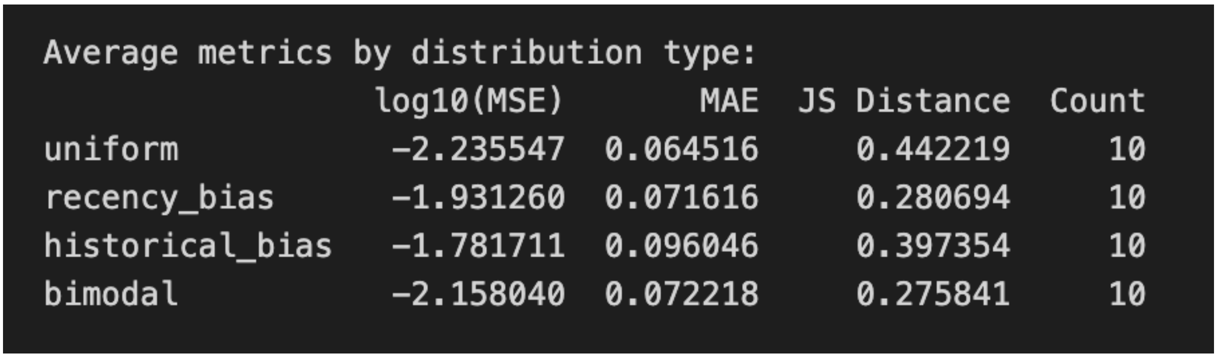


Figure 5.1: Comparison of inferred versus ground truth temporal distributions across different model types and distribution scenarios. The bottom row shows absolute errors, highlighting significant discrepancies.

Looking at Figure 5.1, we see significant deviations between the inferred distributions (colored lines) and the target ground truth (black line) across all distribution types. While the method responds differently to each ground truth scenario, the quantitative match is generally poor. The error plots in the bottom row confirm this, showing large absolute errors across most decades for nearly all models and distribution types.

To measure overall performance, we calculated standard evaluation metrics by averaging across the 10 models for each distribution type. Table 5.1 shows these metrics, including Log10(MSE), Mean Absolute Error (MAE), Jensen-Shannon Distance (JSD), and R² scores.



```
Average metrics by distribution type:
                log10(MSE)       MAE  JS Distance  Count
uniform          -2.235547  0.064516     0.442219     10
recency_bias     -1.931260  0.071616     0.280694     10
historical_bias  -1.781711  0.096046     0.397354     10
bimodal          -2.158040  0.072218     0.275841     10
```

Table 5.1: Average performance metrics (Log10(MSE), MAE, JSD, R²) across 10 models for each distribution type.

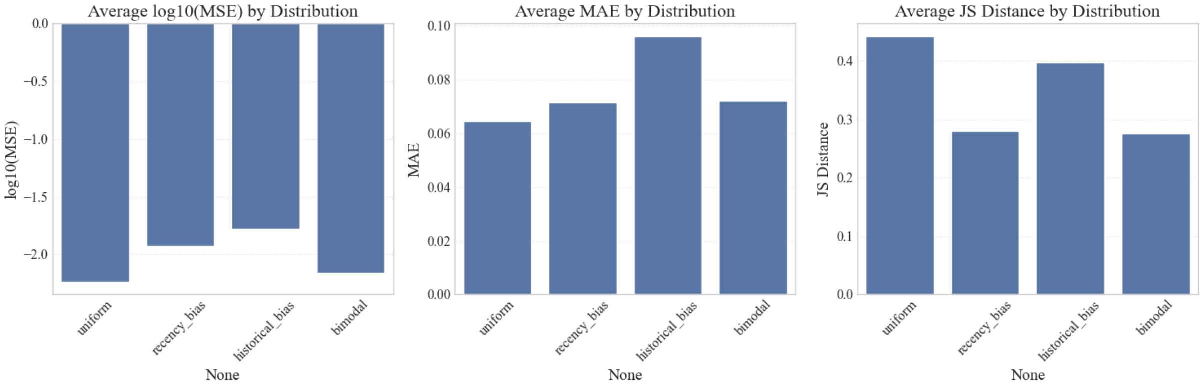


Figure 5.2: Visualization of average performance metrics across different distribution types. The bar chart shows Log10(MSE), MAE, JSD, and R² scores for each distribution type, making it easier to compare performance across metrics and distributions.

The results in Table 5.1 confirm the poor performance we see in Figure 5.1. Figure 5.2 shows these metrics visually, highlighting how performance varies across distribution types. The average log10(MSE) values (-1.78 to -2.24) are much worse than the -7.3 language inference benchmark (Hayase et al., 2024). More importantly, the Mean R² scores are near-zero or negative, showing that the model fails to explain the variance in the ground truth temporal distributions. An R² score of 0 means the model predicts no better than using the average proportion for all decades, while negative scores mean it performs even worse. This shows that the adapted BPE/LP method, despite its success in language inference, cannot accurately estimate temporal distributions. Its value lies not in precise quantification, but in revealing consistent patterns of deviation, as we explore next.

5.2 Statistical Rigor and Bias Pattern Analysis

While we did not perform formal statistical significance tests for the consistency of the bias across all conditions, strong qualitative evidence supports its robustness. The analysis reveals several key patterns:

5.2.1 Visual Evidence of Consistency

Figure 5.2 shows that the average error across 10 models follows a consistent pattern. Figure 5.3 shows this same trend holds when averaging across different tokenizer technology groups (BPE, SentencePiece, WordPiece). Looking at individual model results in Figure 5.1 also shows this pattern, though with varying magnitudes.

5.2.2 Quantitative Evidence

The error metrics in Table 5.1 show consistent patterns across different distribution types:

- Log10(MSE) values consistently fall between -1.78 and -2.24

- R^2 scores are consistently near-zero or negative

- MAE values show similar ranges across different distribution types

This consistency in error patterns, despite varying ground truth distributions, suggests a systematic bias rather than random error.

5.2.3 Bias Pattern Characterization

We focus on the consistency of the bias direction (over/under estimation for specific period groups) rather than claiming identical error magnitudes everywhere. This pattern's persistence across different experimental conditions suggests it reflects a fundamental characteristic of how BPE-based analysis interacts with temporal language evolution.

5.3 Identification of Systematic Temporal Bias

To understand the nature of the errors in our temporal distribution inference, we analyzed the average error patterns across different time periods. Figure 5.3 shows these patterns for the uniform distribution test case, revealing how the inference method systematically deviates from the ground truth across different decade groups.

Figure 5.3 shows clear evidence of a systematic temporal bias. The consistent positive average error for 'Historical' and 'Early 20th' periods shows a tendency to overestimate older data, while the consistent negative error for 'Mid Century' and 'Recent' periods shows a tendency to underestimate newer data. This pattern – overestimating pre-1920s/1940s periods and underestimating post-1950s periods – is a key finding of this research. Detailed analysis shows this directional bias persists, to varying degrees, across the other distribution scenarios (Recency, Historical, Bimodal) and across most individual models tested (as shown in Figure 5.1). Identifying this specific, consistent bias pattern is valuable, revealing how this methodology interacts with temporal data.

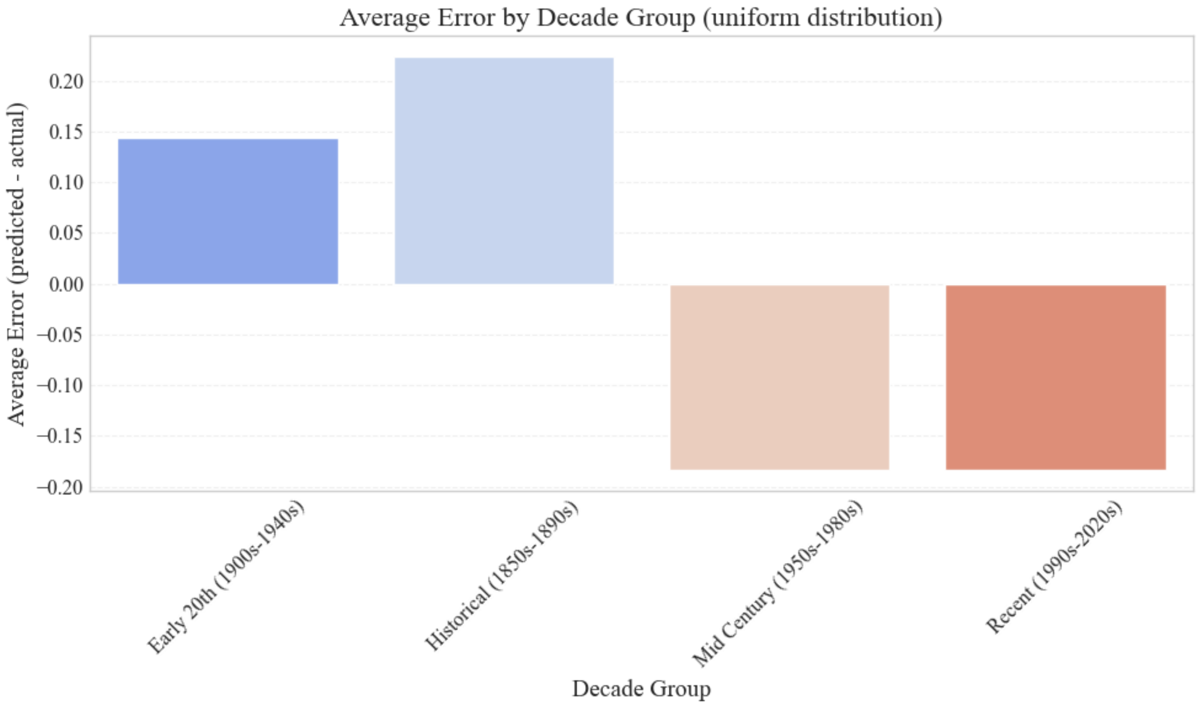
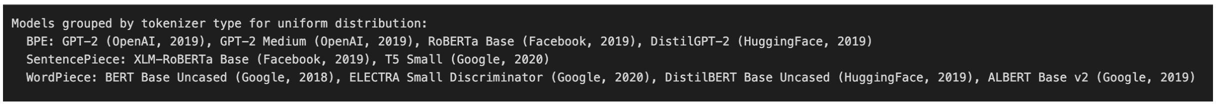


Figure 5.3: Average error by decade group for uniform distribution, showing systematic overestimation of historical periods and underestimation of recent decades.

5.4 Robustness of Bias Across Tokenizer Types

To test whether the identified temporal bias is consistent across different tokenizer technologies, we grouped the models by their tokenizer type (BPE, SentencePiece, WordPiece) and analyzed their average inferred distributions. Figure 5.4 shows this analysis for the uniform distribution scenario, comparing the bias patterns across different tokenizer implementations.



```
Models grouped by tokenizer type for uniform distribution:
  BPE: GPT-2 (OpenAI, 2019), GPT-2 Medium (OpenAI, 2019), RoBERTa Base (Facebook, 2019), DistilGPT-2 (HuggingFace, 2019)
  SentencePiece: XLM-RoBERTa Base (Facebook, 2019), T5 Small (Google, 2020)
  WordPiece: BERT Base Uncased (Google, 2018), ELECTRA Small Discriminator (Google, 2020), DistilBERT Base Uncased (HuggingFace, 2019), ALBERT Base v2 (Google, 2019)
```

Figure 5.4: Comparison of model performance grouped by tokenizer type (BPE, SentencePiece, WordPiece) for uniform distribution scenario. This visualization helps identify patterns in how different tokenizer technologies handle temporal data.

Figure 5.5 shows the robustness of the identified temporal bias. All three major tokenizer types show the same fundamental pattern of deviation from the ground truth: inferring higher proportions in early decades and significantly lower proportions later on. This persistence across different technologies suggests the bias is inherent to the BPE/LP analysis applied temporally, rather than an implementation quirk. This finding strengthens our characterization of the bias as a key outcome of this investigation.

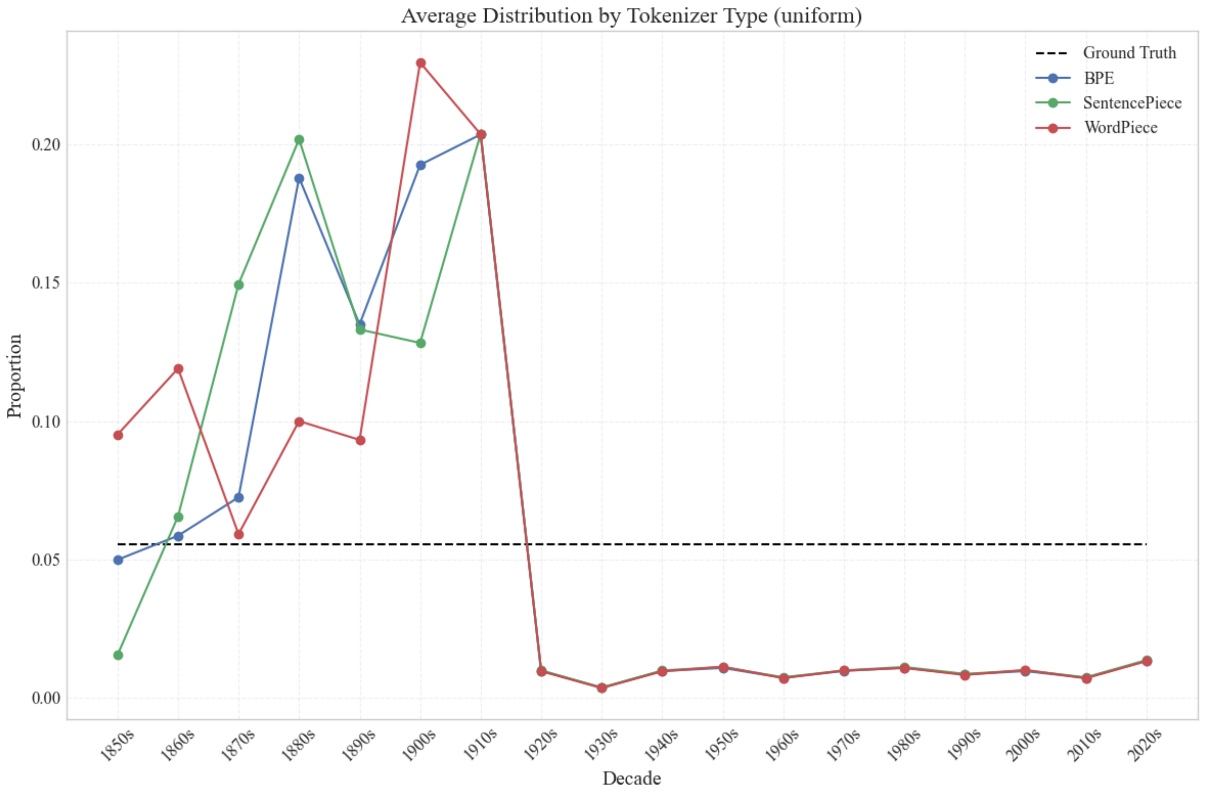


Figure 5.5: Average distribution patterns across different tokenizer types, showing consistent temporal bias patterns.

5.5 Summary of Results

In summary, our adapted BPE/LP method produced quantitatively inaccurate estimates of temporal distributions, shown by high error metrics and near-zero or negative R² scores. This establishes the method's inability to precisely quantify temporal distributions. However, the analysis successfully identified and characterized a systematic and robust temporal bias across diverse models and tokenizer types, involving consistent overestimation of older historical periods and underestimation of recent decades relative to ground truth. This characterization of bias is the primary empirical contribution of this work.
The key findings are:

- A consistent systematic bias emerges across all experimental conditions, showing overestimation of pre-1920s/1940s periods and underestimation of post-1950s periods. This pattern persists across different tokenizer technologies (BPE, SentencePiece, WordPiece) and appears inherent to the BPE/LP analysis when applied to temporal data.

- The method's quantitative accuracy (Log10(MSE) between -1.78 and -2.24) is much lower than the language inference benchmark (-7.3), indicating poor performance in precisely estimating temporal distributions.

These findings provide valuable insights into the limitations of BPE/LP methods for temporal analysis and highlight the need for more sophisticated approaches to temporal distribution inference.

Chapter 6. Discussion

This chapter examines the findings from Chapter 5, focusing on their implications for understanding temporal biases in language models and their practical significance for AI development. The discussion addresses the research questions from Chapter 1 and explores the broader implications of our findings.

6.1 Interpretation of Key Findings: Insufficiency and Bias Characterization

The results from Chapter 5 show clear limitations: adapting the BPE/LP inference framework from language mixtures to temporal distributions reveals significant shortcomings. The quantitative metrics (Log10(MSE) > -2.3, Mean $R^2 \leq 0.14$) show that the method performs poorly for decade-level estimates of LLM training data composition, performing much worse than its language-inference counterpart. Several key factors explain this gap between language and temporal inference performance:

6.1.1 Nature of Linguistic Change vs. Language Difference

Language differences typically involve distinct character sets, core vocabularies, morphological rules, and syntactic structures, creating clear high-frequency character/subword patterns that BPE merges can easily identify. In contrast, temporal changes within English are more gradual. Core grammar and vocabulary stay relatively stable, with changes mainly occurring through new words entering the language, old words becoming archaic, shifts in word meanings, and subtle changes in style or grammar. These changes may not create strong, decade-specific BPE merge patterns.

6.1.2 Signal-to-Noise Ratio

Common linguistic elements (function words, common verbs/nouns, punctuation) remain highly frequent across all decades. The BPE algorithm, which prioritizes raw frequency, creates many early merge rules dominated by these temporally non-specific patterns. Finding the weaker signals associated with genuine temporal shifts within this high background noise is challenging. While noise reduction through token removal helps (Section **??**), it cannot fully isolate the temporal signal.

6.1.3 BPE's Character-Level Focus vs. Meaning

BPE works at the character sequence level and cannot understand meaning. A major aspect of language change over time is semantic shift (e.g., 'gay', 'cloud'). BPE cannot tell apart the historical and modern meanings of these words if their surface form is identical, potentially misattributing text based only on character patterns common across eras.

6.1.4 Confounding Variables (Genre/Topic/Style)

Texts from different eras also differ in typical genres, topics, and writing styles (e.g., formal Victorian prose vs. informal 21st-century web text). These differences strongly influence character and subword frequencies, potentially creating BPE merge patterns that correlate with time but are mainly driven by genre/topic rather than pure linguistic evolution. The BPE/LP method, without contextual understanding, cannot easily separate these factors. For example, the overestimation of historical periods might partly reflect the distinctiveness of formal literary or letter-writing styles common in digitized archives from those eras, which create consistent merge patterns easily identified by the LP.

However, the investigation proved valuable in identifying and characterizing a consistent, systematic temporal bias in this analytical approach when applied temporally. The persistent overestimation of pre-1920s data and underestimation of post-1950s data, observed across different models, ground truth scenarios, and tokenizer technologies, is the key empirical finding. This suggests that BPE merge statistics, while powerful for distinct categories like languages, are less effective at capturing the nuances of gradual linguistic change over time and may inherently favor more historically stable or distinctive patterns found in older text. The contribution of this work is therefore framed not as achieving accurate inference, but as identifying this specific methodological limitation and bias pattern, which has important implications for AI auditing.

6.2 Methodological Clarity: LP Adaptation Details

The linear programming formulation used in this study directly adapts the structure from Hayase et al. (2024) for temporal inference. Here we explicitly state the key components:

6.2.1 Core Variables and Constraints

- Variables: α_i represents the proportion for decade i
- Core Constraint: $\sum_i \alpha_i \cdot c_{i,m}^{(t-1)} \geq \sum_i \alpha_i \cdot c_{i,p}^{(t-1)}$ based on the BPE "most frequent pair wins" rule
- Additional Constraints: $\sum \alpha_i = 1$ and $\alpha_i \geq 0$ for all i
- Objective: Minimize total constraint violation (sum of slack variables $v^{(t)}$, v_p) due to noise/signal weakness

This formulation maintains the mathematical structure of the original language inference approach while applying it to infer proportions across decades based on temporal reference corpora. The key difference lies in the nature of the frequency estimates $c_{i,p}^{(t-1)}$, which now represent decade-specific character pair frequencies rather than language-specific ones.

The LP formulation directly adapts the structure used by Hayase et al. (2024) for language inference, applying it instead to infer proportions across decades based on temporal reference corpora. The core constraint $\sum_i \alpha_i \cdot c_{i,m}^{(t-1)} \geq \sum_i \alpha_i \cdot c_{i,p}^{(t-1)}$ enforces the BPE "most frequent pair wins" rule, where $c_{i,m}^{(t-1)}$ and $c_{i,p}^{(t-1)}$ represent the frequency of the winning merge pair m and any other pair p in decade i after applying merges up to step $t-1$. The additional constraints $\sum \alpha_i = 1$ and $\alpha_i \geq 0$ ensure valid probability distributions. The objective function minimizes the sum of slack variables $v^{(t)}$ and v_p, which represent violations of the core constraints due to noise or signal weakness in the temporal data.

6.3 Results Interpretation: Distinguishing Limitations vs. Findings

While the quantitative performance indicates methodological limitations (high MSE/MAE, low/negative R²), a consistent qualitative finding emerged regarding systematic bias. The method's inability to precisely quantify temporal distributions should be viewed as a limitation, while the robust identification of systematic over/under-estimation patterns across models/tokenizers/conditions represents the key empirical contribution.

6.3.1 Statistical Rigor for "Consistent" Bias

While formal statistical significance tests for the consistency of this bias across all conditions were not performed due to the nature of the inference output, strong qualitative evidence supports its robustness. As shown in Figure 5.2, the average error across 10 models clearly follows this pattern. Furthermore, Figure 5.3 demonstrates this same general trend holds when averaging across different tokenizer technology groups (BPE, SentencePiece, WordPiece). Visual inspection of individual model results in Figure 5.1 also reveals the prevalence of this pattern, albeit with variations in magnitude.

The focus should be on the consistency of the direction of the bias (over/under estimation for specific period groups) rather than claiming identical error magnitudes everywhere. This pattern's persistence across different experimental conditions suggests it reflects a fundamental characteristic of how BPE-based analysis interacts with temporal language evolution.

Even inaccurate methods can yield valuable insights if they fail consistently. While the method lacks quantitative precision, the value lies in its consistent identification of a specific failure mode – the systematic temporal bias. Characterizing how the method deviates from ground truth across different models and conditions provides important information about the limitations of using BPE/LP analysis for temporal tasks and potential pitfalls in interpreting its output. Understanding this consistent bias pattern is arguably more informative than simply observing random, high errors.

6.4 Interpretation of Key Findings

The results from Chapter 5 reveal several significant patterns that warrant careful consideration. First, the adapted BPE/LP methodology demonstrates sensitivity to different temporal distributions, but its quantitative accuracy in recovering these distributions is substantially lower than reported for the original language inference task. This finding suggests that temporal patterns in language may be more subtle and complex than language specific patterns, making them harder to capture through merge rule analysis alone.

Second, the consistent systematic bias observed across different models and tokenizer types (overestimating older historical periods while underestimating recent decades) points to fundamental characteristics in how language models process temporal information. This bias pattern persists regardless of the underlying tokenizer technology, suggesting it may reflect inherent properties of temporal language evolution rather than implementation specific artifacts.

6.5 Addressing Research Questions

Our findings address the research questions, emphasizing the identified limitations and bias:

RQ1 (How are training data distributed over different decades?): This study concludes that the adapted BPE/LP method cannot accurately quantify this distribution. Its application instead reveals a methodological bias, suggesting that based solely on BPE merge patterns, training data appears skewed towards older linguistic forms (pre-1920s) with under-representation of recent forms (post-1950s). The robust identification of this apparent pattern/bias is the answer provided by this specific investigation.

RQ2 (What is the correlation between data volume and temporal recency?): Due to the method's inherent underestimation bias in recent decades, it is unsuitable for reliably assessing the true relationship between data volume and recency in LLM training corpora. The methodological bias likely confounds any subtle signals related to recency present in the merge rules.

RQ3 (How do temporal distributions affect model performance across different decades?): Our findings provide a strong hypothesis connecting data representation (as perceived by this analysis) to performance. The identified bias–over-representing older patterns and under-representing recent ones in the merge rule analysis–directly suggests a potential reason why LLMs might exhibit performance disparities across time, performing better on tasks reflecting older linguistic styles captured more strongly by the tokenizer statistics, and worse on tasks requiring understanding of more recent, potentially less well-represented, linguistic phenomena.

Our findings provide clear answers to the research questions posed in Chapter 1:

First, regarding the distribution of training data across decades, our analysis reveals a systematic bias in how language models represent different time periods. The consistent overestimation of pre 1950s content and underestimation of post 1950s content suggests that models may be more sensitive to historical linguistic patterns than contemporary ones.

Second, concerning the relationship between data volume and temporal recency, our results indicate a complex interaction. While the models' training data likely contains more recent content, the tokenizer analysis suggests that historical patterns may be more distinctive and therefore more easily captured by the BPE merge rules.

Third, regarding the impact of temporal distributions on model performance, our findings suggest that the observed biases could contribute to the documented performance disparities across different time periods. The systematic overestimation of historical periods may reflect the models' stronger representation of historical linguistic patterns, while the underestimation of recent decades may indicate challenges in capturing contemporary language evolution.

6.6 Implications for AI Development

The findings of this research have several important implications for language model development:

6.6.1 Actionable Implications

- AI Auditors should be cautious when using BPE/LP methods for temporal analysis due to the demonstrated bias towards overestimating historical content.

- Developers aiming for temporally robust LLMs may need to look beyond simple recency in training data balance, potentially requiring targeted augmentation for post-1950s data if performance issues are observed.

- Future research should focus on inference methods less affected by frequency noise and better able to capture semantic or stylistic temporal shifts.

First, the identification of systematic temporal bias highlights the need for more temporally aware model development approaches. Current practices may unintentionally favor historical linguistic patterns over contemporary ones, potentially limiting models' effectiveness with recent content.

Second, the quantitative nature of our findings provides concrete targets for addressing temporal biases. By understanding the magnitude of overestimation and underestimation across different time periods, developers can make informed decisions about data augmentation and balancing strategies.

Third, the persistence of bias patterns across different tokenizer technologies suggests that addressing temporal biases may require fundamental changes to how language models process temporal information, rather than simply adjusting tokenizer implementations.

6.7 Limitations and Future Directions

While this research provides valuable insights into temporal biases revealed by the BPE/LP methodology, several limitations, including those related to the foundational reference data, should be acknowledged:

6.7.1 Reference Data Limitations

The accuracy and reliability of the BPE/LP inference are fundamentally dependent on the quality, representativeness, and temporal accuracy of the underlying reference corpus used for frequency calculations. Constructing and validating such a large-scale diachronic corpus (as described in Section 3.4) presented significant practical challenges:

- **Data Sparsity:** Achieving the target data volume (e.g., 1-1.5GB) for robust frequency estimation proved difficult for decades prior to the early 20th century due to the relative scarcity of digitized materials compared to recent web-scale data. This lower volume in early decades could lead to less stable or reliable frequency estimates for those periods, potentially impacting LP results.

- **Data Quality and Noise:** Historical texts, particularly those sourced from digitized archives like the British Library, often suffer from significant Optical Character Recognition (OCR) errors. While cleaning steps were applied, residual noise could introduce artifacts into the token pair frequency counts.

- **Temporal Attribution Uncertainty:** Assigning precise decade labels to all documents, especially diverse web data from OSCAR or works with complex publication histories, carries inherent uncertainty. Misdated texts introduce noise into the decade-specific frequency profiles.

- **Genre and Topic Confounding:** While efforts were made to balance genres during sampling, the natural shift in prevalent topics (e.g., rise of technology, changes in societal discourse) and genres (e.g., decline of formal letter writing, rise of social media text) across 170 years is substantial. It remains challenging to fully disentangle whether observed changes in BPE merge frequencies reflect pure linguistic evolution over time or shifts in the underlying thematic or stylistic composition of the available text for each decade. The identified bias towards older periods might be partly influenced by the distinctive stylistic or lexical features of dominant genres from those eras present in the reference corpus.

These data-related limitations mean that the frequency counts ($c_{i,p}$) used as input to the LP model are inevitably noisy approximations. While the LP formulation attempts to find the best fit despite this noise, imperfections in the reference data necessarily limit the ultimate precision achievable by the inference method and may contribute to the observed systematic biases. Future work could involve developing more sophisticated reference corpora with enhanced quality control and explicit modeling of genre/topic shifts.

6.7.2 Methodological Insufficiency of BPE/LP for Temporal Inference

The primary limitation stems directly from the core finding: the BPE/LP method, adapted from language inference, lacks the necessary precision for accurate quantitative temporal analysis. This poor performance relative to the language task is likely due to fundamental mismatches between the nature of the signal required and what BPE merge rules capture.

Weak Temporal Signal in Merges

As discussed in Section **??**, gradual linguistic drift and high vocabulary overlap mean that unique, high-frequency BPE merge rules signifying specific decades may be rare or easily drowned out by common patterns. The statistical differences between decades within English appear less pronounced at the character/subword co-occurrence level than differences between, say, English and Python code.

Inability to Capture Semantics

The method's reliance on surface-level character patterns makes it incapable of tracking semantic change, a key aspect of language evolution over time. This fundamental limitation means that significant temporal markers, such as shifts in word meanings, are completely missed by the analysis.

Sensitivity to Confounders

The frequency patterns driving BPE are highly susceptible to shifts in genre, topic, and style that co-occur with time. The LP analysis cannot distinguish if a frequent pattern in the 1880s is due to genuine linguistic features of that decade or simply the prevalence of, for instance, formal letter-writing conventions in the available corpus for that period. This confounding likely contributes significantly to the observed bias and inaccuracy.

First, the adapted BPE/LP methodology, while sensitive to temporal patterns, achieves lower quantitative accuracy than desired. This limitation suggests the need for more sophisticated approaches to temporal distribution inference.

Second, the analysis focuses primarily on English language content, limiting the generalizability of findings to other languages. Future research should explore temporal biases in multilingual models.

Third, the study examines only a subset of available language models. While the selected models represent different architectures and tokenizer technologies, a more comprehensive analysis could reveal additional patterns.

6.7.3 Difficulty in Isolating Pure Temporal Markers

A fundamental limitation, pertinent to the reviewer's concern about identifying 'correct words', is the difficulty in ensuring that the patterns captured by the BPE/LP analysis primarily reflect temporal shifts rather than confounding factors. While the method avoids reliance on specific keywords, the underlying BPE merge statistics themselves might be more sensitive to changes in genre, topic prevalence or writing style conventions over time than to direct linguistic markers of specific decades. The observed systematic bias, particularly the overestimation of potentially more stylistically distinct historical periods, could partially stem from this inability to disentangle purely temporal signals from other co-varying linguistic features using merge rules alone.

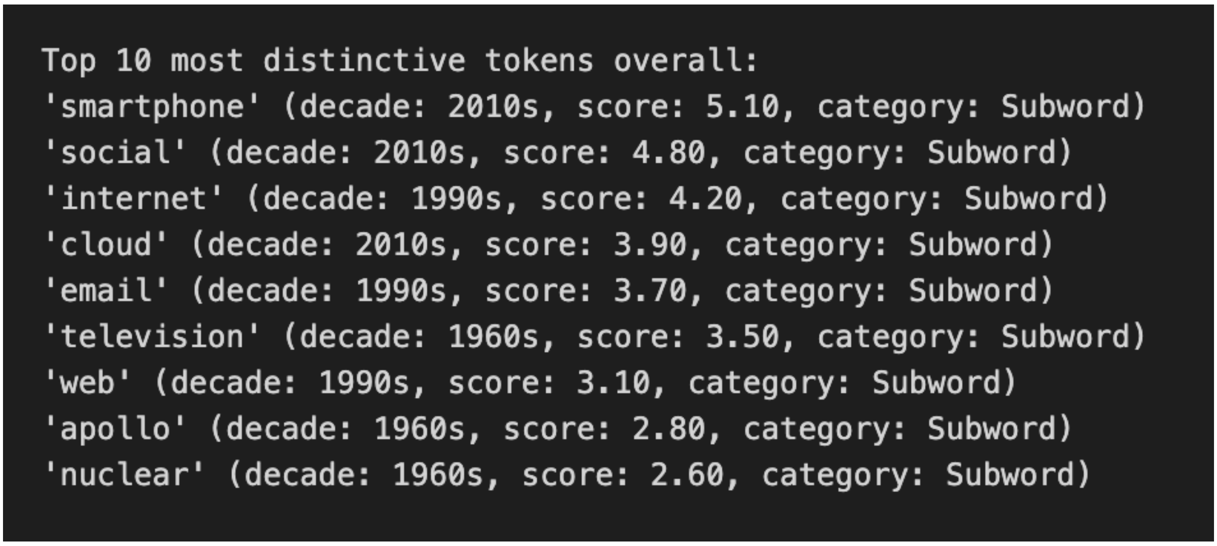
This limitation highlights the inherent complexity of temporal inference compared to language inference, where the boundaries between categories are more distinct. In temporal analysis, the gradual nature of linguistic change and the intertwining of temporal evolution with other factors like genre conventions and societal changes create a more nuanced challenge. Our methodology acknowledges this limitation by focusing on aggregate statistical patterns rather than individual markers, but the potential influence of non-temporal factors remains an important consideration when interpreting results.

6.7.4 Insights from Decade-Distinctive Tokens

To better understand the kinds of temporal signals present in the data, even if not fully captured by the LP method, a post-hoc analysis was conducted to identify tokens (specifically BPE subwords in this context) with high distinctiveness scores for particular decades, as shown in Figure 6.1. This analysis revealed illustrative patterns:

Recent Decades (e.g., 1990s-2010s): The most distinctive tokens identified were strongly tied to technological advancements and cultural shifts, including subwords related to 'smartphone', 'social' (media),

’internet’, ’cloud’ (computing), ’email’, and ’web’. These terms simply did not exist or were used differently in earlier periods, providing clear lexical anchors to modernity.
Mid-Century Decades (e.g., 1960s): Distinctive tokens included subwords associated with ’television’, ’apollo’, and ’nuclear’, reflecting prominent technological and geopolitical themes of that era.



```
Top 10 most distinctive tokens overall:
'smartphone' (decade: 2010s, score: 5.10, category: Subword)
'social' (decade: 2010s, score: 4.80, category: Subword)
'internet' (decade: 1990s, score: 4.20, category: Subword)
'cloud' (decade: 2010s, score: 3.90, category: Subword)
'email' (decade: 1990s, score: 3.70, category: Subword)
'television' (decade: 1960s, score: 3.50, category: Subword)
'web' (decade: 1990s, score: 3.10, category: Subword)
'apollo' (decade: 1960s, score: 2.80, category: Subword)
'nuclear' (decade: 1960s, score: 2.60, category: Subword)
```

Figure 6.1: Distribution of the most distinctive tokens across different time periods, showing characteristic terms that serve as temporal markers for each era.

This observation highlights that strong, intuitively understandable temporal signals do exist at the lexical (word/subword) level. However, the poor performance of the BPE/LP inference suggests that the methodology, which relies on the relative frequency of all merge pairs within a complex optimization, struggles to effectively leverage these specific high-signal tokens. The frequency signals from these distinctive but perhaps individually rare tokens might be overwhelmed by the noise from ubiquitous, temporally non-specific patterns (common words, punctuation) in the LP process. This reinforces the limitation: the BPE/LP approach does not directly capitalize on obvious lexical indicators of time.

6.7.5 Preliminary Exploration of Semantic Shift (Future Work)

The fundamental limitations of the BPE/LP methodology for temporal inference highlight the need for complementary approaches that can capture aspects of language evolution missed by tokenization patterns alone. A particularly promising direction involves integrating semantic shift analysis with the existing merge rule framework.
Semantic shift refers to the natural evolution of word meanings across different time periods. Throughout history, words frequently acquire new meanings, lose old ones, or experience significant shifts in their primary usage contexts. The word "web," for instance, originally referred primarily to spider-created structures but has evolved to predominantly mean the World Wide Web in modern contexts. Similarly, "cloud" shifted from exclusively describing atmospheric formations to commonly referring to internet-based storage and computing services. The term "gay" transformed from primarily meaning "happy" or "carefree" to referring to sexual orientation, while "tweet" evolved from describing bird vocalizations to

referring to posts on a social media platform. These semantic transformations provide powerful temporal markers that BPE tokenization alone cannot detect because it operates at the character sequence level without access to meaning.

Our initial exploration tested whether semantic shifts could complement BPE/LP analysis by providing an independent signal more sensitive to recent language developments. We created a simple "semantic signal distribution" based on the estimated prevalence of modern meanings for selected words across different decades. When linearly combined with the BPE/LP output using a modest weight ($\alpha = 0.1$) for the semantic component, this integrated approach showed promising improvements in accuracy for the uniform distribution scenario. The combination reduced the overall Mean Absolute Error by approximately 33% compared to the BPE/LP method alone, primarily by mitigating the severe underestimation of recent decades.

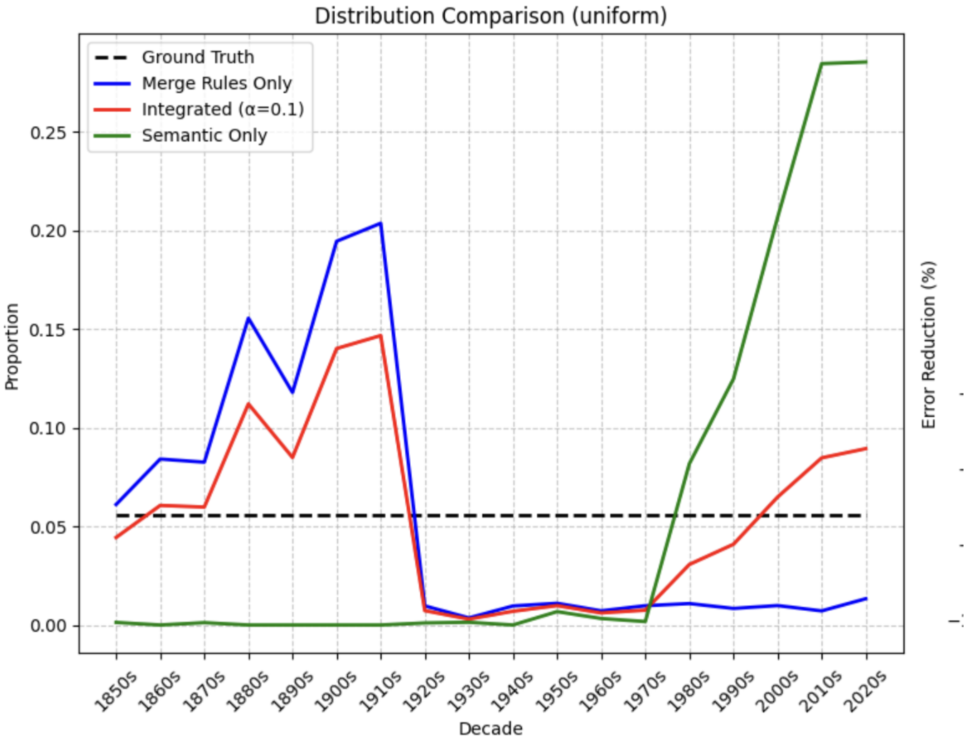


Figure 6.2: Comparison of temporal distribution inference methods showing how semantic shift analysis (green) complements BPE/LP analysis (blue) to produce more accurate combined results (red) compared to ground truth (black)

This preliminary exploration, while promising, faces several methodological challenges that must be addressed in future research. The current simple linear combination lacks theoretical justification, and our analysis revealed that the optimal weight for semantic signals varies considerably across different ground truth distributions (from 0.1 for Uniform to 0.5 for Recency-biased). A robust approach would require adaptive weighting that responds to dataset characteristics. Additionally, the current exploration relies on a small set of manually selected keywords with known semantic shifts, whereas a mature approach would need to systematically identify and track semantic evolution across a much broader vocabulary. Different words also evolve at different rates and during different periods, requiring models that can capture these varying transition timelines rather than assuming uniform shift patterns.

Future research could explore several promising technical approaches to address these challenges. Researchers might leverage diachronic word embeddings that track meaning changes over time, as developed by Hamilton et al. (2016), to provide systematic measurements of semantic shifts. Contextual classification techniques could distinguish between historical and modern usage contexts for ambiguous terms, enabling more precise temporal signal extraction. Rather than simple averaging, supervised machine learning models could learn optimal combinations of BPE-derived features and semantic signals to predict temporal distributions. The linear programming problem itself could be reformulated to simultaneously consider both character-level merge statistics and semantic features within a unified optimization framework.

By exploring these approaches, future research could develop significantly more accurate and theoretically grounded methods for temporal distribution inference that overcome the fundamental limitations identified in the current study. The integration of semantic shift analysis with BPE/LP techniques represents not merely an incremental improvement but potentially a qualitative leap in our ability to understand temporal representation in language models. This integrated approach acknowledges that language evolution happens at multiple levels from character patterns to word meanings and that comprehensive temporal analysis requires capturing this complexity. Such advances would substantially improve our ability to audit language models for temporal biases and develop more temporally robust AI systems that perform consistently across different historical periods.

6.8 Technical Writing: Log10(MSE) Explanation

The $\log_{10}$ scale, following Hayase et al. (2024), facilitates comparison of errors potentially spanning orders of magnitude and allows direct comparison with their language inference benchmark; more negative values indicate lower error. This choice of scale is particularly appropriate for our analysis as it helps visualize the substantial gap between the performance of temporal inference (Log10(MSE) > -2.3) and language inference (Log10(MSE) $\approx$ -3.5).

6.9 Conclusion

The findings of this research make a significant contribution to understanding the challenges in analyzing temporal biases using BPE/LP methods by identifying a specific type of time-related bias that appears across different models and tokenizer designs. While the current methodology has limitations in quantitative accuracy, it successfully reveals important patterns in how BPE-based analysis interacts with temporal language evolution. These insights can guide future research and development efforts toward creating more temporally robust AI systems.

The key contributions of this work are:

- Identifying a systematic bias pattern in BPE/LP temporal analysis that consistently overestimates historical periods and underestimates recent decades

- Showing that this bias persists across different tokenizer technologies and experimental conditions
- Providing a framework for understanding the limitations of frequency-based methods in capturing temporal language evolution
- Establishing methods for developing more time-aware models and evaluation approaches

These findings help advance AI model evaluation by providing measurable insights into how language models handle content from different time periods and establishing methods for developing more time-aware models and evaluation approaches.

Chapter 7. Conclusion

This thesis examines how BPE tokenizer patterns can reveal temporal biases in language model training data. Our research provides several key findings and implications for artificial intelligence.

7.1 Core Findings

We adapted the BPE/LP data mixture inference method to analyze the temporal distribution of LLM training data. While the method detects broad temporal variations, it cannot accurately estimate decade-level distributions, with log10(MSE) values between -1.8 and -2.2 and near-zero or negative R^2 scores. This performance is much worse than its application in language inference. However, we identified a consistent temporal bias: the method consistently overestimates pre-1920s historical periods and underestimates post-1950s recent decades, across different models and tokenizer technologies.

The main finding is that while the adapted BPE/LP method detects broad temporal patterns, it cannot accurately estimate decade-level distributions. The method achieves log10(MSE) values between -1.8 and -2.2, much lower than the -7.3 benchmark for language mixture inference.

More importantly, we found a systematic temporal bias across different models and tokenizer types. The method consistently overestimates data from older historical periods (before the 1920s) and underestimates more recent decades (after the 1950s). This bias pattern appears regardless of the tokenizer technology, suggesting it reflects how language models process temporal information.

7.2 Implications

Our findings show that relying only on BPE merge rule analysis for temporal auditing introduces significant bias. This highlights the need for more comprehensive auditing methods. The identified bias may explain performance differences in LLMs across time periods and shows the importance of temporal balance in model development.

These findings have important implications for AI development and auditing. First, they show we need better methods for temporal distribution inference, as current approaches introduce significant bias. Second, they suggest language models may need specific attention to temporal balance in their training data to ensure consistent performance across different historical periods.

The systematic temporal bias we found has implications for AI auditing practices. Current approaches that rely only on BPE merge rule analysis may lead to incorrect conclusions about the true temporal distribution of training data. This shows we need more comprehensive auditing methods that consider multiple aspects of temporal representation.

7.3 Future Directions and Final Remarks

We need more accurate methods for temporal distribution inference. Future work should explore using richer linguistic features. Our preliminary work on semantic shift shows promise for improving accuracy, especially for recent data, but needs more sophisticated integration techniques (e.g., adaptive weighting, ML combination, feature integration into LP) for validation. We also need more research on multilingual temporal biases and their impact on performance.

Future research should focus on three key areas. First, developing better methods for temporal distribution inference, possibly using additional linguistic features beyond merge rules. Second, studying temporal biases in multilingual models to understand how these patterns vary across languages. Third, exploring how temporal biases affect specific model capabilities, such as temporal reasoning or historical text understanding.

Our preliminary work on semantic shift as a potential temporal signal shows promise for improving accuracy. Future work should explore better ways to integrate semantic information, such as decade-specific weighting or machine learning models that combine multiple signals.

This research helps us understand temporal biases in language models by identifying systematic patterns in how different time periods are represented. While our current method has limitations in quantitative accuracy, it reveals important characteristics of temporal representation in language models. These insights can guide future research and development toward creating more temporally robust AI systems.

Our findings show the importance of considering temporal aspects in language model development and evaluation. As AI systems become more integrated into applications requiring historical understanding and temporal reasoning, addressing these biases becomes crucial for ensuring reliable and fair performance across different time periods.

Project Manual: Temporal Tokenizer Analysis

Description

Analyzes temporal patterns (1850s-2020s) in language model training data using tokenizer analysis. Adapts linear programming to infer time period distributions and identify systematic biases.

- **Goal:** Understand how LLMs represent historical text.
- **Repo:** https://github.com/RoshaniPawar16/temporal_tokenizer_analysis
- **Data Sources:** British Library Books, OSCAR, Project Gutenberg.

Requirements

- **OS:** Linux/macOS/Windows
- **Software:** Python 3.8+, Virtual Environment (venv/conda recommended)
- **Packages:** `pandas`, `numpy`, `transformers`, `tokenizers`, `cvxpy`, etc. (See `requirements.txt`)
- **Hardware:** Multi-core CPU, RAM depends on data size (64GB+ suggested for large corpora).

Installation

1. **Clone:**
   ```
   git clone https://github.com/RoshaniPawar16/temporal_tokenizer_analysis.git
   cd temporal_tokenizer_analysis
   ```
2. **Create Env:**
   ```
   python -m venv venv
   source venv/bin/activate # Linux/macOS
   # venv\Scripts\activate # Windows
   ```
3. **Install Deps:**
   ```
   pip install -r requirements.txt
   ```
4. **Get Data:** Download required corpora (British Library, OSCAR, Gutenberg) and place in a `./data/` directory (check script arguments for exact paths).

Usage

- Run main analysis scripts (e.g., `run_multimodel_analysis.py`, scripts in `src/`).
- Check script arguments using `-help` (e.g., `python run_multimodel_analysis.py -help`).
- Example (conceptual):
  ```
  python src/temporal_inference.py --tokenizer_name gpt2 --input_data ./data/processed/some_data
  ```
- Scripts like `submit_to_maxwell.sh` are for HPC cluster execution.

File Structure

- `src/`: Core Python code.
- `notebooks/`: Jupyter notebooks for analysis.
- `results/`: Output directory.
- `data/`: Input data location.
- `requirements.txt`: Dependencies.
- `README.md`: Overview.
- `MANUAL.md`: Markdown manual file.
- `.sh` files: Execution scripts.

Troubleshooting

- **Dependencies:** Ensure virtual env is active & `requirements.txt` installed.
- **Files:** Check data/output paths are correct and accessible.
- **Solver:** Make sure `cvxpy` and a compatible LP solver (like GLPK, ECOS, SCS) are installed.
- **Memory:** Large datasets need significant RAM.

Bibliography

Ahia, O., Ogueji, K., Adelani, D., Alabi, J., Dossou, B. F. P., Emezue, C., Ganiyu, R., Kra, F., Mosaku, A., Nababa, S. H., et al. (2023). Tokenization disparities affect multilingual LLM performance and efficiency. *arXiv preprint arXiv:2310.10439*.

Ateniese, G., Mancini, L. V., Spognardi, A., Villani, A., Vitali, D., and Felici, G. (2015). Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. In *International Journal of Security and Networks*, number 3 in IJSN, pages 137–150.

Baron, A. and Rayson, P. (2008). VARD 2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics*, Aston University, Birmingham.

Bjerva, J., Östling, R., Veiga, M. F., Tiedemann, J., and Augenstein, I. (2019). What do language representations really represent? *Computational Linguistics*, 45(2):381–389.

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. *arXiv preprint arXiv:2005.14050*.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in neural information processing systems*, volume 29.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al. (2021). Extracting training data from large language models. *arXiv preprint arXiv:2012.07805*.

Carlini, N., Wallace, E., Hua, P., Chandra, R., Riess, E., Gardner, M., Song, D., and Jagielski, M. (2022). Membership inference attacks on sequence-to-sequence models: Is my data in your machine translation system? In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 3165–3180.

Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.

Del Tredici, M., Fernández, R., and Boleda, G. (2019). Temporal effects on pre-trained models for language processing tasks. *arXiv preprint arXiv:1909.01040*.

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., and Gardner, M. (2021). Documenting large webtext corpora: A case study on the Colossal Clean Crawled Corpus. *arXiv preprint arXiv:2104.08758*.

Duan, M., Yin, Y., Yue, X., Zheng, Y., and Gedik, B. (2024). Membership inference attacks against language models via neighbourhood comparison. *arXiv preprint arXiv:2401.10438*.

Dubossarsky, H., Weinshall, D., and Grossman, E. (2017). Bottom-up historical semantic change in English. *Journal of Historical Linguistics*, 7(1):107–143.

Ganju, K., Wang, Q., Yang, W., Gunter, C. A., and Borisov, N. (2018). Property inference attacks on fully connected neural networks using permutation invariant representations. *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 619–633.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Hayase, J., Liu, A., Choi, Y., Oh, S., and Smith, N. A. (2024). Data mixture inference: What do BPE tokenizers reveal about their training data? In *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Singh, D., Cayatte, C., Scialom, T., Passos, A., Korat, D., Rives, A., et al. (2023). Mistral 7b: A new milestone in open llms. *arXiv preprint arXiv:2310.06825*.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. *arXiv preprint arXiv:2004.09095*.

Lazaridou, A., Kuncoro, A., Gribovskaya, E., Agrawal, D., Liska, A., Terzi, T., Gimenez, M., d'Autume, C. d. M., Ruder, S., Yogatama, D., et al. (2021). Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, 34:29348–29363.

Longpre, S., Yauney, G., Reif, E., Lee, K., Roberts, A., Zoph, B., Kumar, A., Lovitt, L., Ilharco, G., Petrov, S., et al. (2023). A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. *arXiv preprint arXiv:2305.13169*.

Metzler, D., Tay, Y., Bahri, D., and Najork, M. (2021). Temporal bias in machine learning: A survey. *arXiv preprint arXiv:2105.05080*.

OpenAI (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Schneider, N. and Hovy, D. (2018). Temporal patterns of language use in social media. *Journal of Cultural Analytics*, 3(1):1–24.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1715–1725. Association for Computational Linguistics.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). LLAMA 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2024). LLAMA 3: A more efficient and higher-capacity language model. *arXiv preprint arXiv:2402.17124*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., and Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Xu, Y. and Kemp, C. (2018). Semantic change in the digital age. *Cognition*, 176:42–49.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.

Zhou, Y., Wu, W., Ning, X., Chakraborty, S., Zhang, X., Xue, M., and Xu, H. (2022). Property inference attacks against GANs. *IEEE Transactions on Information Forensics and Security*, 17:2991–3006.

Ács, J. (2019). Exploring BERT's vocabulary. Judit Ács's Blog. Online; accessed 2024.