Filling the Temporal Void: Recovering Missing Publication Years in the Project Gutenberg Corpus Using LLMs

Omar Momen, Manuel Schaaf and **Alexander Mehler**
Text Technology Lab
Goethe-Universität Frankfurt am Main
omar.momen.amin@gmail.com, {manuel.schaaf,mehler}@em.uni-frankfurt.de

Abstract

Analyzing texts spanning long periods of time is critical for researchers in historical linguistics and related disciplines. However, publicly available corpora suitable for such analyses are scarce. The Project Gutenberg (PG) corpus presents a significant yet underutilized opportunity in this context, due to the absence of accurate temporal metadata. We take advantage of language models and information retrieval to explore four sources of information – Open Web, Wikipedia, Open Library API, and PG books texts – to add missing temporal metadata to the PG corpus. Through 20 experiments employing state-of-the-art Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) methods, we estimate the production years of all PG books. We curate an enriched metadata repository for the PG corpus and propose a refined version for it, which includes 53 774 books with a total of 3.8 billion tokens in 11 languages, produced between 1600 and 2000. This work provides a new resource for computational linguistics and humanities studies focusing on diachronic analyses. The final dataset and all experiments data are publicly available[1].

1 Introduction

Diachronic Text Analysis (DTA) is a pivotal area of study that encompasses various disciplines and objectives. Linguists examine the evolution of language over time, analyzing changes in word meanings (Giulianelli et al., 2020), morphological structures (Bowern and Evans, 2015), and syntactic patterns (Krielke, 2021). Natural Language Processing (NLP) practitioners engage with texts from different historical periods to develop models that can handle language variation over time (Dhingra et al., 2022; Ren et al., 2023). Humanities scholars study social, political, and cultural trends by analyzing texts from specific time periods (Dinu and Uban, 2023). To facilitate effective DTA studies, it is essential to have access to a publicly available corpus of texts annotated with their time of production. Texts that can only be roughly dated are ultimately worthless to DTA when it comes to performing analyses based on precise time periods.

Although Project Gutenberg (PG)[2] offers a high potential corpus for DTA, it is rarely considered in many studies. PG is a widely recognized digital library that provides access to a large collection of free eBooks. Established in 1971 by Michael S. Hart, PG aims to promote the creation and distribution of digital literature. As of February 2025, PG offers over 75 000 free eBooks in more than 60 languages, covering a diverse range of literary works, including classic novels, historical texts, and reference materials.

Despite its extensive size, linguistic diversity, and historical depth, PG has not been widely adopted as a standard corpus in DTA. A key limitation is the absence of temporal metadata specifying the year of writing or publishing each book in the corpus. The only temporal information available in the PG records is the release date of the book on the PG web portal, which does not correspond to the book's time of writing or publication. This lack of temporal annotation not only undermines the potential of the PG corpus, but also introduces ambiguity and leads to misinterpretation. One reason is that external platforms and web crawlers of online libraries and bibliographic databases often misattribute the PG release date as the publication date of the book, hampering efforts to fill this gap.

This paper addresses the problem of missing temporal metadata in PG books by highlighting its core challenges (§ 3) and exploring potential solutions (§ 4). We conduct a comprehensive set of experiments (§ 5) on all the PG books published before 01.09.2024 (our cut-off date), leveraging a

[1] https://github.com/OmarMomen14/pg-dates

[2] https://gutenberg.org/

Findings of the Association for Computational Linguistics: ACL 2025, pages 17318–17334

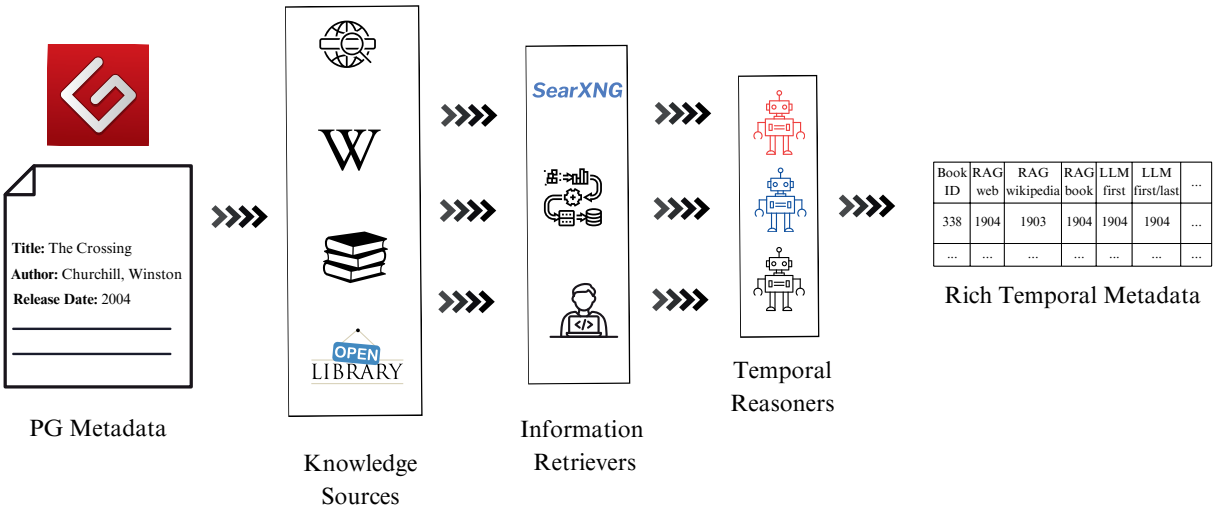

Figure 1: Overview of PG Temporal Labelling Experiments: We use 4 sources of knowledge, Open Web, Wikipedia, PG Books Texts, and Open Library API. We then experiment with 3 methods of information retrieval, a search engine, RAG pipelines, and curated retrieving strategies based on our own observations. We feed the retrieved contents as contexts within prompts to 3 variants of LLMs (Llama, Gemma, and Mistral Nemo). We then retrieve the temporal answer from the LLMs responses and validate them. Eventually, we get 20 temporal estimations for 72 103 PG books, with varying degrees of time accuracy that is measured based on our proposed validation set.

diverse range of sources and methods to infer the missing temporal information. As illustrated in Figure 1, we utilize web pages, knowledge bases, online libraries, and the textual content of the books themselves. We employ search engines, LLMs, and RAG methods to extract and interpret relevant temporal cues from these sources. To the best of our knowledge, this is the first study to systematically integrate such a broad set of sources and methods for enriching the PG corpus with structured temporal annotations. Additionally, we conduct an in-depth analysis of PG records and uncover a subset of books annotated with original publication years. We use this collection as a validation set for evaluating our temporal estimation approaches. The key contributions of this work are:

1. **Comprehensive Temporal Estimates**: We develop 20 different temporal estimates of the production years of 72 103 PG books, using various information sources and ML methods.
2. **Refined Subset with Reliable Labels**: We propose a subset of PG books with reliable year-level production labels, achieving an exact match accuracy of 90% on the validation set and a median error margin of six years.
3. **Retrieved Contents for Temporal Metadata**: We publish all the retrieved contents from the Open Web, Wikipedia, and Open Library that are related to the production time of each PG book.

2 Related Work

Accurate time-labeling of books remains a critical yet underexplored challenge in the study of language change. Researchers investigating diachronic changes in texts usually rely on standard historical corpora, but these corpora are often constrained by issues such as licensing restrictions, coarse temporal granularity, limited time intervals, and narrow genre representation. For example, the Corpus of Historical American English (COHA) (Davies, 2010) contains about 406 million words from 1810–2009 with year labels; but it contains only English texts, and its use is limited by fees. While the Corpus of Contemporary American English (COCA) (Davies, 2008) offers a collection of texts containing about one billion words from 1990–2020, it is also not freely available, and its time interval is not sufficient for many studies. Freely accessible resources such as the Corpus of Late Modern English Texts (CLMET) (Smet, 2005), which includes 34 million words from 1710-1920 with year labels, and the Hansard UK Parliament corpus (Hansard Corpus, 2007), with 1.6 billion words from 1803–2005, provide valuable alternatives, even though they suffer from either limited size or limited genres. Comparable efforts in other languages further underscore these challenges: GerParCor (Abrami et al., 2024) and DeuParl (Walter et al., 2024) compile extensive collections of German parliamentary debates with

detailed metadata that enable fine-grained analyses of formal language change. All these resources highlight the inherent limitations of current corpora and emphasize the need for a large, diverse, freely available corpus with accurate time labels to robustly track language change.

Before the era of LLMs, few approaches have been developed to assign a precise production year to a document. Early methods, such as (Garcia-Fernandez et al., 2011), estimated publication dates by analyzing time-series features derived from a text's content. (Figueira et al., 2017) reframed document dating as a multi-class text classification problem, leveraging Bidirectional Gated Recurrent Units to capture subtle temporal cues. More recently, (Vashishth et al., 2018) propose employing Graph Convolutional Networks to jointly exploit syntactic and temporal structures within documents. Nonetheless, the overall accuracy of these methods remains very modest relative to the precision required for robust DTA, highlighting the need for further improvements in document dating techniques.

To the best of our knowledge, only one prior study has addressed the problem of temporal labelling of the PG books. (Gerlach and Font-Clos, 2020) attempted to standardize PG by annotating the PG books with estimated time labels derived from the author's lifespan. However, this sole study resulted in coarse estimations (author's lifespan) of the production year of a book, rather than an exact year of production. It also failed to find any time estimations for books of unknown authors, or for authors of unknown lifespan.

3 Core Challenges

Determining the exact year when a book was written varies in difficulty from book to book, depending on factors such as the popularity of the book or its author, the number of its editions, or its published metadata in freely accessible sources. The books in the PG vary drastically in this regard.

In literary records, the most commonly documented temporal attribute of a book is its date of publication, rather than its date of creation. However, a book may be republished multiple times, resulting in different publication dates in different editions. Consequently, it is difficult to determine a single authoritative year for a book's publication, let alone its original creation date. In practice, the earliest known year of publication is often used as an approximation of the creation date. To avoid confusion between the creation year and the publication year of a book, we consider the earliest known publication year of a book and call it the book's *production year*.

Estimating the production year of a book is particularly challenging due to the scarcity and fragmentation of relevant information. For well-known books or authors, this information can be inferred from the book's metadata in open access sources, and it can also be validated by evidence from biographies or literary studies. However, such sources are limited and often unavailable for lesser-known works. Moreover, the available records are often inconsistent, incomplete, or even contradictory. The process of aggregating and reconciling these scattered data points is non-trivial, especially when dealing with large collections spanning multiple genres and languages.

To further analyze the task of determining a book's production year, identify its challenges, and gain insight into potential solutions, we conducted a preliminary study: we randomly selected 15 book titles (10 in English and 5 in German) from the PG corpus and attempted to determine their production years. The primary goal was to assess the availability and reliability of temporal information in different sources. To maximize coverage, we leveraged a variety of information sources known to contain bibliographic and historical data. Specifically, we examined **Online Libraries** represented by the Library of Congress, Google Books, WorldCat, and Open Library, **Knowledge Bases** represented by Wikipedia, **LLMs with Web Search** represented by ChatGPT-4o, and the **Book's Content** itself by skimming the first and last pages of the book.

By manually querying these sources, we evaluated their effectiveness in retrieving accurate production years and highlight the key challenges associated with our task. The detailed annotations and queries/prompts used in this experiment are presented in Appendix A. According to this preliminary study, we find that online libraries often lack information about books, and in many cases only provide the same metadata available in PG, which unfortunately adds a confusing layer to the problem. Wikipedia generally contains fairly accurate information, but only for well-known books or authors who have their own pages. ChatGPT-4o demonstrates great capability in identifying accurate production years by leveraging its web search capabilities alongside its reasoning capabilites. A

Source	# Successful determination
LoC Catalog	5
WorldCat	7
Open Library	8
Google Books	9
Wikipedia	8
ChatGPT	13
Book Content	12

Table 1: Results of the preliminary study: A successful determination of the publication year based on a source (row) is only considered as such if the year attested by the source corresponds to the earliest year of publication identified across all sources.

key insight from this experiment is that, in most cases, the content of the book itself provides clues about its production year. Such clues are typically found in either the first few pages or the last few pages of the book. Table 1 presents the number of books for which we successfully identified the production year from each source.

From this preliminary study, we conclude that no source alone provides accurate temporal information about a book's publication year. In order to obtain an accurate dating, the results from different sources should be aggregated and heuristics should be applied to select the most likely correct result.

4 Methods

Figure 1 outlines our approach; in this section, we elaborate and justify it. We discuss the three main layers in our methodology: knowledge sources, information retrievers and temporal reasoners.

Knowledge Sources Based on our preliminary study, we decide to use four knowledge sources in our temporal labelling experiments: **(1) The Open Web**, since temporal metadata can be found in numerous types of web pages (e.g., online libraries, online bookstores, genre-specific book platforms); **(2) Wikipedia**, since it contains accurate data in dedicated pages for popular authors and books; **(3) Open Library API**, since it contains structured data about millions of books; and **(4) The Book Content**, since according to our preliminary study, accurate temporal information about production years are often mentioned in the books pages.

Information Retrievers To retrieve the most relevant information regarding the production year of a book from the open web, we use Searxng[3] an open-source **Search Engine** that aggregates the search results of more than 80 search engines including Google search. Additionally, we use **RAG Pipelines** to retrieve the most relevant texts from the Wikipedia dumps and the Books Contents. Finally, we use our **Own Observations** from our preliminary study that the usual locations of production year mentions in PG books are on their first and last pages.

Temporal Reasoners The retrieved content for a given book – which is not structured data – needs to be reasonably processed to get a valid estimate of its production year. A human can easily do such processing, but we are interested in processing the whole PG corpus. We decide to use state-of-the-art (SOTA) LLMs, which have good temporal reasoning capabilities. Studies such as (Chu et al., 2024; Qiu et al., 2024) show a significant lag between humans and LLMs in many temporal reasoning skills, however, they show satisfactory performance when asked to identify the oldest or most recent year in a passage. We use variants of Llama3.1 (Meta, 2024), Gemma2 (Google, 2024), and Mistral Nemo (Mistral AI team, 2024).

Validation Method A central result of our preliminary study is the discovery of metadata fields in the biography tables on the PG web portal that are not included in the metadata files provided by the maintainers. First, for books uploaded to the PG web portal after January 2022, there is a new field, *Original Publication*, which contains information about the publisher and publication year of the book. Second, we observe that most books contain a recently added field with an automatically generated summary. These summaries can provide an approximate period in which the book was written, such as "early 19th century" or "late 18th century". Third, we identify a metadata field recording the *Reading Ease Score*, which quantifies the linguistic complexity of the book. These metadata provide valuable information for solving our detection task, so we collect the summaries and the readability scores to investigate their relevance to our study. We extract the original publication years for books uploaded after January 2022, resulting in a subset of 7 168 books, which represents about 10% of the entire PG collection at our cutoff date. We use this subset as a validation set in our study.

[3] https://docs.searxng.org/

Experiment	Val.	Acc.	Err.
Baselines			
$\text{Base}_{\text{llama}}$	.65	.09	13y
$\text{Base}_{\text{gemma}}$	.77	.09	11y
$\text{Base}_{\text{mistral}}$	.99	.10	14y
Open Web			
LLM_{web}	.72	.78	9y
RAG			
$\text{RAG}_{\text{wikipedia}}$	.93	.29	20y
RAG_{book}	.99	.57	10y
Context Based On Observations			
$\text{LLM}_{\text{first}}$	.95	.65	9y
$\text{LLM}_{\text{first/last}}$	**.99**	**.84**	**6y**
$\text{LLM}_{\text{first/last/summary}}$	.99	.83	7y
$\text{LLM}_{\text{whole}}$	.71	.78	9y
$\text{LLM}_{\text{first/last}}$ with different prompts			
$\text{LLM}_{\text{detailed-prompt}}$	.98	.70	24y
$\text{LLM}_{\text{lifespan}}$	.99	.79	8y
$\text{LLM}_{\text{first/last}}$ with different models			
Llama	.99	.81	7y
$\text{Llama}_{\text{full precision}}$	.99	.77	10y
$\text{Llama}_{\text{70B}}$	.99	.82	8y
Mistral Nemo	.98	.84	7y
WizardLM	.94	.66	26y
Online Queries			
Open Library API	.76	.53	17y
Coarse Estimations			
Author Lifespan (33s)	.84	.63	–
Author Lifespan (100s)	.84	.79	–
Summary (33s)	.99	.81	–
Summary (100s)	.99	.91	–
$\text{LLM}_{\text{first/last}}$ (10s)	.99	.90	–
$\text{LLM}_{\text{first/last}}$ (20s)	.99	.92	–
$\text{LLM}_{\text{first/last}}$ (33s)	.99	.95	–
$\text{LLM}_{\text{first/last}}$ (50s)	.99	.96	–
$\text{LLM}_{\text{first/last}}$ (100s)	.99	.97	–

Table 2: Evaluating all experiments on the validation set (7 168 books). **Val.** is the percentage of *valid* responses (where we could find a match for a valid year pattern in the response). **Acc.** is the exact matching *accuracy*. **Err.** is the median of the year difference for all non-matching valid answers. The LLM used in all experiments is *gemma-2-9b-it*, unless otherwise stated. All LLMs are 8-bit quantized. Exact matching is measured by finding all patterns of *YYYY* in the LLM response, and considering the oldest of them as the LLM answer; however, in most cases there's only one *YYYY* pattern found in the response.

5 Experiments

We evaluate several variations of our approach using a strict metric that measures the exact match between the estimated production year and the year annotated in our validation set. Results of all experiments are presented in Table 2, with corresponding queries and prompts detailed in Appendix C.

5.1 Baseline Experiments

As a baseline, we prompt an LLM to estimate the production year of each book based on its pretrained knowledge, without additional context. Since PG books are often included in large-scale LLM training data, this serves as a reference for evaluating subsequent methods. We use *Llama3.1-8B-Instruct*, *Gemma-2-9b-it*, and *Mistral-Nemo-Instruct-2407* as examples of SOTA LLMs. The baselines perform poorly, indicating that pretrained knowledge alone is unreliable for this task and that additional context is necessary. Although we explicitly instruct the models not to answer if uncertain (Appendix C), they frequently generate incorrect responses, particularly $\text{Base}_{\text{mistral}}$. For the remaining experiments, we report results only for the Gemma model (unless otherwise noted), as it consistently outperforms the other two models.

5.2 Open Web Experiment

We implement this approach by searching for the production year of each PG book on the open web. Using Searxng, which retrieves snippets directly answering the search query, we feed these snippets into an LLM and prompt it to infer the production year based solely on this context. LLM_{web} significantly outperforms the baselines. Missing .28 answers result from either no web results being found or the LLM responding with *no answer*.

5.3 RAG Experiments

We perform two RAG experiments using different sources as knowledge bases. We follow the same RAG procedure as in (Semnani et al., 2023).

Wikipedia as Knowledge Base We conduct an experiment using Wikipedia as a structured knowledge base for an LLM. (Semnani et al., 2023) chunk and embed the Wikipedia dump (01.08.2024) in ten languages using the BAAI General Embedding (BGE) model *bge-m3*[4] and index the embeddings with the Qdrant[5] vector database. For each PG title, we query Qdrant for the most relevant content related to the book's production year and retrieve the top three chunks. We then prompt an LLM to extract and interpret the estimated production year from these chunks. $\text{RAG}_{\text{wikipedia}}$ improves over the baseline, but we notice that the chunks often confuse the LLM, leading to larger errors (20y) compared to the baseline (11y).

Book Content as Knowledge Base In this experiment, we apply the same RAG procedure (chunking, embedding and indexing) on the texts of the

[4] https://huggingface.co/BAAI/bge-m3
[5] https://github.com/qdrant/qdrant

PG books, we then retrieve the top three relevant chunks from each book that may reference the production year. RAG_{book} performs significantly better than $\text{RAG}_{\text{wikipedia}}$ but remains less effective than LLM_{web}, raising questions about the retrieval capabilities of the RAG pipelines.

5.4 Context Based On Observations Experiments

Our preliminary study indicates that temporal information is most often found in a book's first or last pages. To leverage this insight, we design experiments exploring different manual retrieval strategies, model configurations, and prompts.

First Pages as Context In this experiment, $\text{LLM}_{\text{first}}$, we provide an LLM with the first 10 000 tokens of each book and prompt it to estimate the production year based solely on this context. The model finds answers for .95 of the books, with .65 matching the labelled year in the validation set. These promising results motivate further exploration.

First and Last Pages as Context We extend the previous experiment by including both the first 5 000 and last 5 000 tokens as context to assess whether information at the end of a book improves prediction accuracy. $\text{LLM}_{\text{first/last}}$ achieves the best results across all our experiments, reinforcing our observation that production years are often mentioned in a book's first or last pages.

First & Last Pages with Summary as Context To refine the estimation, we enrich the context with an automatically generated book summary, which often provides an approximate writing era (e.g., "early 19th century") and may help constrain the model's inference. However, $\text{LLM}_{\text{first/last/summary}}$ does not improve over $\text{LLM}_{\text{first/last}}$, instead showing a slight drop in accuracy and a higher error margin. This suggests that additional information may introduce confusion rather than help for LLM.

Whole Book In Sequence To account for the possibility that temporal information appears mid-book, we design an iterative experiment. We first provide the LLM with 10 000 tokens and prompt it to determine if this context is sufficient to infer the production year. If more content is needed, we sequentially supply the next 10 000 tokens until an answer is found or the full text is processed. Contrary to expectations, $\text{LLM}_{\text{whole}}$ does not outperform $\text{LLM}_{\text{first/last}}$, likely due to early stopping before reaching more relevant information at the book's end.

Prompt Variations Since prompt wording affects LLM predictions, we augment the best-performing experiment ($\text{LLM}_{\text{first/last}}$) with prompt variations. In $\text{LLM}_{\text{detailed-prompt}}$, we provide explicit instructions on inferring the production year. While in $\text{LLM}_{\text{lifespan}}$, we add information about the author's birth year and death year (when available), and guide the model to consider this information as a reference period (see Appendix C). However, neither variant improves over the original prompt.

Model Variations We evaluate multiple LLMs of different sizes using the $\text{LLM}_{\text{first/last}}$ setting to assess their effectiveness. Alongside the baseline models, we test *Llama3.1-8B* (full precision), *Llama3.1-70B*, and *WizardLM-2-8x22B*. Results indicate that models in the 7–9 billion parameter range are sufficient for this task, as larger models do not yield significant improvements.

5.5 Querying an Online Library

We query the Open Library API for production years. This method improves over the baseline, but it can not compete with the experiments that utilize prompting LLMs with supplied context. We notice that many titles lack publication year metadata, and some are entirely absent. This underscores the limitations of relying on external bibliographic sources for estimating production years.

5.6 Error Analysis

To identify factors contributing to incorrect predictions, we conduct a feature analysis on books (of the validation set) that consistently lead to errors. We examine correlations between incorrect estimates and various features, aiming to develop criteria for confidence levels on estimates of the full corpus. Specifically, we analyze the effects of the predicted *century*, book *language*, *length*, the *frequency of year patterns* in the extracted context, and whether the *predicted year explicitly appears in the provided context*. Most features show no significant difference between correct and incorrect predictions. However, **explicit mention of the predicted year in the context** strongly correlates with accuracy. Our analysis reveals that: (1) 83% of books have their predicted year explicitly mentioned in the provided context. (2) Within the validation set, 90% of books containing the predicted year in their context yield correct estimates. These

findings suggest that filtering out books where the predicted year is absent from the extracted text could improve the corpus's overall accuracy.

5.7 Alternative Coarse Estimations

Since predicting an exact year remains challenging (with our best-performing method achieving 84% accuracy), we explore alternative estimation strategies using broader time ranges to refine and validate our predictions.

Using the Author's Lifespan Since a book must be written within the author's lifespan, we retrieve and process all available author metadata from PG. For books with known authors, we estimate the production year as the midpoint of their lifespan and refine this estimate by mapping it to broader time ranges of 10, 20, 33, 50, and 100 years.

Using Automatically Generated Summaries The automatically generated book summaries indicate an approximate writing era. We extract these estimates and map them to numerical ranges, e.g., "Early 19th century" → 1800–1833, "Mid-19th century" → 1834–1866, and "Late 19th century" → 1867–1900. Finally, we generalize these estimates to century-level ranges (100 years).

5.8 Results Discussion

The results of the RAG experiments contradict our expectations, likely due to the challenges in RAG retrieval capabilities (Barnett et al., 2024; Martin et al., 2024). In contrast, experiments using manually selected book content as LLM context yield strong results, particularly when leveraging the first and last few pages, aligning with our preliminary study.

To quantify our contribution to PG temporal labelling, we compare the coarse estimation of $\text{LLM}_{\text{first/last}}$ (33s) to the only existing estimate in the literature, Author Lifespan (33s). Our approach finds estimates for both known and unknown authors, covering .99 of PG books. It achieves a matching accuracy of .95, significantly outperforming the Author Lifespan method, which reaches only .63.

6 Extended PG Corpus

We release the results of all experiments for 72 103 PG books, along with the cleaned raw text used as context. For each book, we provide a tokenized version and a sentence-split version. This dataset serves as a valuable resource for DTA studies and research on LLMs' temporal reasoning capabilities.

6.1 Metadata

Our experiments generate extensive temporal metadata for all PG books up to the cutoff date. We also enhance the original PG metadata by incorporating structured information from the PG web interface. The original metadata includes only Book ID, Title, Author Name, Language(s), PG Release Date, and Browsing Category. To improve dataset utility, we enrich it with: **(1) Rich Temporal Estimations**: 20 year-level estimations plus 15 coarser estimations in 10, 20, 33, 50, and 100-year levels. **(2) Author's Lifespan**: Retrieved (when available) to constrain possible writing periods. **(3) Book Summary**: Automatically generated summaries from PG bibliographic pages. **(4) Structural Statistics**: Number of characters, tokens, and sentences per book. **(5) Genre Classification**: Mapping 64 browsing categories into 8 general genres. **(6) Readability Score**: A numerical complexity measure from PG bibliographic pages.

6.2 Corpus Statistics

The enriched PG corpus comprises 72 103 books across 68 languages, totalling 5.3 billion tokens and 255 million sentences, making it a valuable resource for linguistic evolution studies. The most accurate temporal estimation, based on validation set evaluation, comes from the $\text{LLM}_{\text{first/last}}$ experiment. This method determines the production year for 99.9% of PG books up to the cutoff date, achieving 84% exact-year match accuracy on the validation set, with a median error of 6 years for the remaining 16%.

6.3 Filtering and Heuristics

To enhance the reliability of temporal estimations and refine the dataset for DTA studies, we apply heuristics and filters based on our observations. These constraints prioritize accuracy and consistency. A full list of filters and heuristics is provided in Appendix D. Applying these refinements results in a corpus of 53 774 books, totalling 3.8 billion tokens and 182 million sentences. The dataset remains multilingual, covering 11 languages. Detailed corpus statistics, including distributions across languages, genres, and estimated writing periods, are provided in Appendix B.

6.4 Three Demonstrative Use Cases

To illustrate the applicability and limitations of the refined PG corpus, we analyze temporal trends of key features. We visualize the evolution of book length, sentence length, and readability based on our most reliable year-level estimations, and also based on the traditional author's lifespan estimation (33-year range). Figure 2 shows the average book length (tokens per book) per decade from 1600–2000 for both estimations. Fluctuations in the 17th century likely stem from the limited number of books, leading to unstable estimates. A clear upward trend appears in the 18th and 19th centuries, followed by a decline in the 20th century.

Figure 3 visualizes sentence length over time. Except for periods with sparse data (e.g., the 16th century and late 20th century), a downward trend is observed, suggesting progressively shorter sentences. Figure 4 tracks readability via the Ease Score, where higher values indicate greater readability. Apart from fluctuations due to data sparsity, the trend suggests that books have become increasingly easier to read over time. These analyses highlight the potential of the extended PG corpus for studying long-term trends while emphasizing the impact of data sparsity in certain periods.

A Wilcoxon signed-rank test was conducted to evaluate the impact of year-level estimations on measured attributes. The test reveals a significant difference in sentence length between author lifespan estimations and year-level estimations, $p = .023$. Similarly, a significant difference was found in ease scores between the two estimation methods, $p = .030$. However, the test revealed no significant difference in book lengths between the two estimation methods, $p = .468$. Thus, for two of the variables in our test, we find significant differences that cannot be overlooked by corresponding analyses.

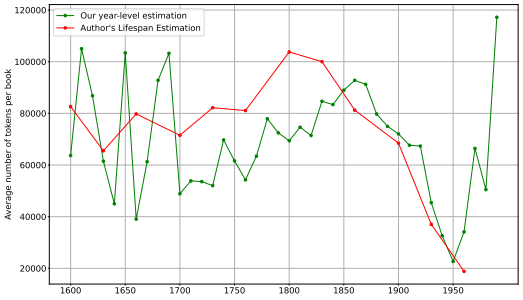

Figure 2: Evolution of books length from 1600 to 2000. For the green line, each point x represents books produced in the closed interval $[x, x+9]$ as per our year-level estimation. While for the red line, each point x represents books produced in the closed interval $[x, x+33]$ as per author's lifespan estimation estimation.

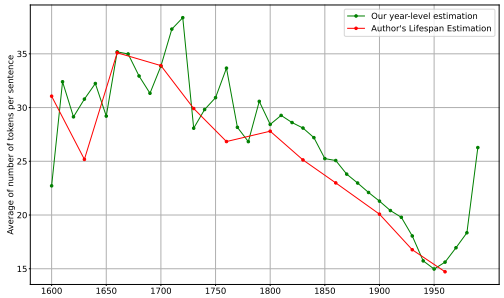

Figure 3: Evolution of sentence length from 1600 to 2000. For the green line, each point x represents books produced in the closed interval $[x, x+9]$ as per our year-level estimation. While for the red line, each point x represents books produced in the closed interval $[x, x+33]$ as per author's lifespan estimation estimation.

Conclusion

We highlight the need for a publicly available corpus of diverse texts written over a long span of time with accurate production year labels. We systematically show the difficulty of retrieving these year labels for books in the PG corpus. We explore a wide range of sources and methods backed by the reasoning capabilities of SOTA. Eventually, a rich metadata dataset including multi-level temporal information for the PG corpus is curated. We also refine the corpus by applying careful heuristics and filters resulting in a 3.8-billion tokens corpus with reliable production year labels that achieve 90% accuracy on the validation set. This work introduces valuable assets for further research in computational linguistics and humanities studies.

Limitation

By recognizing these limitations, we aim to provide a transparent account of the potential constraints of our study and to inform future research endeavours seeking to build upon our work.

Imbalance in Temporal and Linguistic Distribution The PG corpus exhibits significant disparities in the representation of different time periods and languages. Certain eras and languages are underrepresented, leading to an uneven distribution

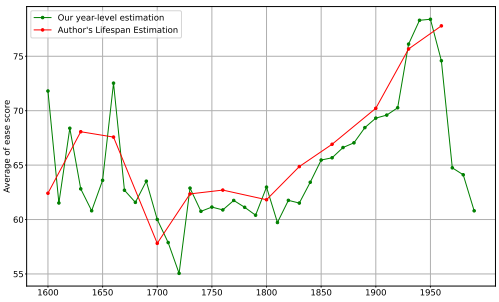

Figure 4: Evolution of readability score from 1600 to 2000. For the green line, each point x represents books produced in the closed interval $[x, x+9]$ as per our year-level estimation. While for the red line, each point x represents books produced in the closed interval $[x, x+33]$ as per author's lifespan estimation estimation.

of data. This imbalance may affect the robustness and applicability of any linguistic analyses.

Uncertainty in the Validation Set The validation set employed in our experiments is constructed based on available metadata and inferred temporal information. However, the accuracy and reliability of this set are contingent upon the correctness of the source data and the methods used for inference. Any inaccuracies or inconsistencies in the original metadata or the inference process could introduce uncertainty into the validation set.

Acknowledgments

This work is carried out within the BIOfid project. BIOfid is supported by the Deutsche Forschungsgemeinschaft (DFG). We would also like to acknowledge the efforts made by the maintainers of PG, without their efforts to build and maintain this great open-access repository, our study would never have existed. We specifically thank Greg Newby and Johannes Seikowsky for being responsive and cooperative regarding our questions and inquiries about the corpus and its metadata. We also thank the anonymous ARR reviewers for their valuable suggestions. Finally, we acknowledge the support of the project "CRC 1646: Linguistic creativity in communication" and Prof. Dr. Sina Zarrieß for making this work presented in ACL 2025.

References

Giuseppe Abrami, Mevlüt Bagci, and Alexander Mehler. 2024. German parliamentary corpus (GerParCor) reloaded. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7707–7716, Torino, Italia. ELRA and ICCL.

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. 2024. Seven failure points when engineering a retrieval augmented generation system. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI*, CAIN '24, page 194–199, New York, NY, USA. Association for Computing Machinery.

C. Bowern and B. Evans. 2015. *The Routledge Handbook of Historical Linguistics*. Routledge Handbooks in Linguistics. Taylor & Francis.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2024. TimeBench: A comprehensive evaluation of temporal reasoning abilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1204–1228, Bangkok, Thailand. Association for Computational Linguistics.

Mark Davies. 2008. The corpus of contemporary American English: 450 million words, 1990-present.

Mark Davies. 2010. The corpus of historical American English: 400 million words, 1810–2009.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Liviu P. Dinu and Ana Sabina Uban. 2023. Analyzing stylistic variation across different political regimes. In *Computational Linguistics and Intelligent Text Processing*, pages 110–123, Cham. Springer Nature Switzerland.

Rui Figueira, Daniel Gomes, and Bruno Martins. 2017. Automatic dating of textual documents.

Anne Garcia-Fernandez, Anne-Laure Ligozat, Marco Dinarelli, and Delphine Bernhard. 2011. When was it written? automatically determining publication dates. In *String Processing and Information Retrieval*, pages 221–236, Berlin, Heidelberg. Springer Berlin Heidelberg.

Martin Gerlach and Francesc Font-Clos. 2020. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy*, 22(1):126.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change

with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Google. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Hansard Corpus. 2007. Hansard UK Parliament Corpus. https://www.english-corpora.org/hansard/. Accessed: 12 February 2025.

Marie-Pauline Krielke. 2021. Relativizers as markers of grammatical complexity: A diachronic, cross-register study of english and german. *Bergen Language and Linguistics Studies*, 11(1):91–120.

Andreas Martin, Hans Friedrich Witschel, Maximilian Mandl, and Mona Stockhecke. 2024. Semantic verification in large language model-based retrieval augmented generation. *Proceedings of the AAAI Symposium Series*, 3(1):188–192.

Meta. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Mistral AI team. 2024. Mistral Nemo. Accessed: 2024.

Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo Ponti, and Shay Cohen. 2024. Are large language model temporally grounded? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7064–7083, Mexico City, Mexico. Association for Computational Linguistics.

Han Ren, Hai Wang, Yajie Zhao, and Yafeng Ren. 2023. Time-aware language modeling for historical text dating. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13646–13656, Singapore. Association for Computational Linguistics.

Sina Semnani, Violet Yao, Heidi Zhang, and Monica Lam. 2023. WikiChat: Stopping the hallucination of large language model chatbots by few-shot grounding on Wikipedia. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2387–2413, Singapore. Association for Computational Linguistics.

Hendrik Smet. 2005. The Corpus of Late Modern English Texts. *ICAME-Journal*, 29.

Shikhar Vashishth, Shib Sankar Dasgupta, Swayambhu Nath Ray, and Partha Talukdar. 2018. Dating documents using graph convolution networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1605–1615, Melbourne, Australia. Association for Computational Linguistics.

Tobias Walter, Celina Kirschner, Steffen Eger, Goran Glavaš, Anne Lauscher, and Simone Paolo Ponzetto. 2024. Diachronic analysis of german parliamentary proceedings: Ideological shifts through the lens of political biases. In *Proceedings of the 2021 ACM/IEEE Joint Conference on Digital Libraries*, JCDL '21, page 51–60. IEEE Press.

A Manual Investigation Results

In Tables 5– 17, we present our annotations for the preliminary study that we performed in § 3

PG Metadata	
ID	65220
Title	The Answer
Author(s)	Emil Petaja, W. E. Terry [Illustrator]
Language	English
Sources	
Book Content	**1951**. The year is spotted on the first page of the book content.
LoC Catalog	**N/A**. No records are found for the book.
Google Books	**N/A**. No records are found for the book.
Wikipedia	**1951**. No page for the book, only the author's Wikipedia page (in German) mentions the book and dates it to 1951.
Open Library	**N/A**. No records are found for the book; the author's page lists some works, but this one is not on the list.
WorldCat	**2021**. One record is found, with a publication year of 2021, which is the time of releasing the book on PG.
ChatGPT	**1951**. It also provides the source of this information, which is a database of fiction books[6].
Comments	The book is a short story published in a magazine in 1951. No other specific publication dates are found. The magazine accepts open submissions from authors, so it could have been written earlier.

Table 3: Results for Book ID 65220.

PG Metadata	
ID	58237
Title	Motion pictures, 1940-1949
Author(s)	Library of Congress. Copyright Office
Language	English
Sources	
Book	**1953**. It is found on the first page of the book.
Content	
LoC	**N/A**. Surprisingly, no record is found for the book in the Library of Congress catalog.
Catalog	
Google	**1953**. One record is found with the publication date under the "Published" field.
Books	
Wikipedia	**N/A**. No page for the book.
Open	**1953**. One record is found showing the publication date under "Publish Date".
Library	
WorldCat	**1953**. One record is found showing the publication date under "Year".
ChatGPT	**1953**. It mentioned the source is the Internet Archive.
Comments	The book is a catalog from the Library of Congress covering motion pictures from 1940 to 1949. It is a reference book and not well-known.

Table 4: Details for Book ID 58237.

PG Metadata	
ID	73247
Title	Old Harmless
Author(s)	Roy Norton, 1869-1942
Language	English
Sources	
Book	**1920**. It is found on the last page of the book
Content	content, where it mentions that the story appeared in the December 7, 1920 issue of "The Popular Magazine".
LoC	**N/A**. No record is found for the book.
Catalog	
Google	**N/A**. No record is found for the book.
Books	
Wikipedia	**N/A**. No page is found for the book or the author.
Open	**N/A**. No record is found for the book; a few
Library	records are found for the author, but not for this specific book.
WorldCat	**N/A**. No record is found for the book; many books are listed for authors with the same name, but not this book.
ChatGPT	**1920**. The source is the book text itself.
Comments	This item is a short story published in a magazine in 1920. It is not a well-known book.

Table 5: Details for Book ID 73247.

PG Metadata	
ID	70626
Title	The Tragedy of Monomoy Beach
Author(s)	Clarkson P. Bearse, 1871-1952
Language	English
Sources	
Book	**1943**. Found in the metadata on the record
Content	page and in the first page of the book, showing the copyright date as 1943.
LoC	**N/A**. No record is found for the book.
Catalog	
Google	**1943, 2023**. Two records are found with
Books	different years.
Wikipedia	**N/A**. No page is found for the book or the author.
Open	**1943**. One record is found showing the year
Library	under "Publish Date".
WorldCat	**1943, 2023**. Two records are found with different years.
ChatGPT	**1943**. The source is the PG book and the Archive.org record.
Comments	The book is a historical account of Monomoy Beach in Cape Cod, focusing on shipwrecks. The correct date of 1943 is confirmed in multiple sources.

Table 6: Details for Book ID 70626.

PG Metadata	
ID	38056
Title	Abraham Lincoln and the London Punch
Author(s)	William S. Walsh, 1854-1919 [Editor]
Language	English
Sources	
Book	**1909**. Found in the first page of the book,
Content	showing the publication date as March 1909.
LoC	**1909**. One record is found showing the pub-
Catalog	lishing date under "Published/Created".
Google	**1909, 2018**. Multiple records are found with
Books	different publishing years.
Wikipedia	**N/A**. No page is found for the book; a page is found for another person with the same name of the author.
Open	**1909**. One record is found showing the pub-
Library	lishing year under "Publish Date".
WorldCat	**1909, 2011**. Multiple records are found with different publishing years.
ChatGPT	**1909**. The sources are the Open Library record and the onlinebooks website.
Comments	The book is a collection of cartoons, comments, and poems about Abraham Lincoln. There are variations of the book title across different sources, but the content remains the same. The correct date, 1909, is confirmed in multiple sources.

Table 7: Details for Book ID 38056.

PG Metadata	
ID	34856
Title	My Own Story
Author(s)	Emmeline Pankhurst, 1858-1928
Language	English
Sources	
Book Content	**1914**. Found in the first pages of the book, showing the publication date as 1914.
LoC Catalog	**1914**. Two records are found, both showing the publication date as 1914.
Google Books	**1914**. Multiple records are found, most showing recent publication dates (2000s), but the publication date (1914) is noted under "Original published".
Wikipedia	**1914**. No page is found for the book; the author's page mentions the book and its date as of 1914.
Open Library	**1914, 2000s**. Multiple records are found, some showing the publication date as 1914 under "Publish Date", others showing 2000s.
WorldCat	**1914, 2000s**. Multiple records are found, some showing the publication date as 1914 under "Year", others showing 2000s.
ChatGPT	**1914**. The source is the Archive.org record.
Comments	The book is an autobiography of Emmeline Pankhurst. The correct date of 1914 is confirmed in multiple sources. Open Library was particularly helpful for dates and records. The book is quite popular.

Table 8: Details for Book ID 34856.

PG Metadata	
ID	63931
Title	Guest Expert
Author(s)	Allen Kim Lang, 1928-; Paul Orban, 1896-1974 [Illustrator]
Language	English
Sources	
Book Content	**1951**. Found in the first page of the book, with a reference to its publication in *Planet Stories* January 1951.
LoC Catalog	**N/A**. No record is found for the book.
Google Books	**2020**. One record is found for the PG release of the book.
Wikipedia	**N/A**. No page is found for the book or the author.
Open Library	**N/A**. No record is found for the book.
WorldCat	**2020**. One record is found for the PG release of the book.
ChatGPT	**1951**. The source is isfdb.org
Comments	The book is a short science fiction story published in *Planet Stories* in 1951. It is not a well-known book, and its date was challenging to find across multiple sources. ChatGPT performed well in identifying the relevant date.

Table 9: Details for Book ID 63931.

PG Metadata	
ID	14639
Title	Punch, or the London Charivari, Volume 152, February 28, 1917
Author(s)	Various
Language	English
Sources	
Book Content	**1917**. Found in the first page of the PG book, and the volume date is also mentioned in the title.
LoC Catalog	**N/A**. No record is found for the book.
Google Books	**2005**. Only the PG release date is found.
Wikipedia	**N/A**. Only a page for the magazine is found; no specific page for this volume.
Open Library	**N/A**. No record is found for this volume; other volumes of the magazine are listed.
WorldCat	**2005**. Only the PG release date is found.
ChatGPT	**1917**. The source is Archive.org and PG records.
Comments	The book is a volume of the humor and political magazine *Punch*, which published monthly volumes from 1841 to 1996. Many volumes of the magazine are available in the PG catalog. The correct date is included in the PG title.

Table 10: Details for Book ID 14639.

PG Metadata	
ID	12409
Title	The Story of the Philippines
Author(s)	Murat Halstead, 1829-1908
Language	English
Sources	
Book Content	**1898**. Found in the first page of the book, with a publication date of 1898.
LoC Catalog	**1898**. One record is found showing the publishing date under "Published/Created".
Google Books	**1898**. One record is found showing the publishing date under "Published" and "Originally published".
Wikipedia	**1898**. The author's page mentions the book and its publication date of 1898.
Open Library	**1898**. Multiple records are found, most showing the publication date as 1898 under "Publish Date".
WorldCat	**1898**. One record is found showing the publishing date under "Publisher".
ChatGPT	**1898**. The source is Archive.org and the PG book.
Comments	The book is a history of the Philippines and the Spanish-American War. It is a well-known book, and the correct date of 1898 is confirmed in most sources.

Table 11: Details for Book ID 12409.

PG Metadata	
ID	32950
Title	Camp and Trail
Author(s)	Stewart Edward White, 1873-1946; Fernand Lungren, 1857-1932 [Illustrator]
Language	English
Sources	
Book Content	**1906**. Found in the first pages of the book, showing the copyright date of 1906.
LoC Catalog	**1906, 1911**. Two records are found: one with the publication date as 1906 (in the notes field), another with 1911.
Google Books	**1907**. One record is found showing the publishing date under the field "Originally published".
Wikipedia	**1907**. The author's page mentions the book and its date of 1907.
Open Library	**1907**. Multiple records are found, most showing the publication date as 1907 under "Publish Date", some showing 1997.
WorldCat	**2019**. A new edition of the book is found with the publication date of 2019.
ChatGPT	**1907**. The source is Archive.org and the PG book.
Comments	The book is somewhat popular and was copyrighted in 1906, 1907.

Table 12: Details for Book ID 32950.

PG Metadata	
ID	30289
Title	Nach Amerika! Ein Volksbuch. Sechster Band
Author(s)	Friedrich Gerstäcker, 1816-1872; Karl Reinhardt, 1818-1877 [Illustrator]
Language	German
Sources	
Book Content	**1855**. Found in the first page of the book, with the publication date of 1855.
LoC Catalog	**N/A**. No record is found for the book.
Google Books	**1855**. One record is found with the PG release date of 2009; another record shows 1855 under "Originally published".
Wikipedia	**N/A**. No page is found for the book; the author's page does not mention it.
Open Library	**N/A**. No record is found for the book.
WorldCat	**2009**. Multiple records are found, most showing the publication date as 2009.
ChatGPT	**1855**. The source is Manybooks.net and the PG book itself.
Comments	The book is somewhat popular. The date of 1855 is confirmed in multiple sources, including Manybooks.net, which appears to be a useful resource for finding original publication dates of some PG books.

Table 13: Details for Book ID 30289.

PG Metadata	
ID	55026
Title	Sämmtliche Werke 3: Abende auf dem Gutshof bei Dikanka; Phantastische Novellen
Author(s)	Nikolai Vasilevich Gogol, 1809-1852; B. Schenrock [Commentator]; Otto Buek, 1873-1966 [Editor]; Frieda Ichak, 1879-1952 [Translator]; Alexandra Ramm, 1883-1963 [Translator]; Ludwig Rubiner, 1881-1920 [Translator]
Language	German
Sources	
Book Content	**1910**. Found in the first page of the book, with the publication date of 1910.
LoC Catalog	**N/A**. No record is found for the book.
Google Books	**1831, 1910, 2017**. Multiple records are found: most records show 1910 as the publication date; one record mentions 1831 as the original publication date.
Wikipedia	**1831**. A trick page for the book under a different name, mentions 1831 as the publication date.
Open Library	**N/A**. No record is found for the book.
WorldCat	**1910**. Multiple records are found, most showing the publication date as 1910.
ChatGPT	**1831**. The source is Wikipedia, pointing to 1831 as the original publication date.
Comments	The book was published as part of Gogol's collected works in 1910, though the original story dates back to 1831. This example highlights the challenges in tracking publication dates when multiple editions exist.

Table 14: Details for Book ID 55026.

PG Metadata	
ID	14142
Title	Land und Volk in Afrika, Berichte aus den Jahren 1865-1870
Author(s)	Gerhard Rohlfs, 1831-1896
Language	German
Sources	
Book Content	**1870**. Found in the first page of the book, with a publication date of 1870.
LoC Catalog	**N/A**. No record is found for the book.
Google Books	**1870**. Multiple records are found, most showing 1870 as the publication date.
Wikipedia	**1870**. The author's page mentions the book and its publication date of 1870.
Open Library	**1884**. One record is found showing 1884 as the publication date.
WorldCat	**1870**. Multiple records are found, with many showing 1870 as the publication date.
ChatGPT	**1884, 1870**. The source is the Open Library record, though another source (Thalia) suggests 1870.
Comments	The book is a travel account of Africa, published in 1870. It has multiple editions, and while some sources suggest 1884, the oldest found edition is from 1870, which is the preferable date.

Table 15: Details for Book ID 14142.

PG Metadata	
ID	14915
Title	Das Nibelungenlied
Author(s)	Unknown (Translated by Karl Simrock)
Language	German
Sources	
Book Content	**N/A**. No specific date could be determined from the book content.
LoC Catalog	**1881**. Multiple records are found, with one showing a publication in Stuttgart in 1881.
Google Books	**1839, 1994, 1997, 2005**. Various editions are found with different dates.
Wikipedia	**1200, 1827**. The original poem was written in 1200, and the first printed translation was in 1827.
Open Library	**Multiple dates**. Various editions are found with different dates.
WorldCat	**1867**. Multiple records are found, most showing 1867 as the publication date.
ChatGPT	**1200, 1827**. Suggested 1200 for the original poem, and 1827 for the first printed translation.
Comments	The original poem, *Das Nibelungenlied*, was written in medieval German around 1200. The version on PG is a translation by Karl Simrock, first published in 1827, and the specific PG version is likely from 1868. The book has many editions, translations, and is a well-known epic poem.

Table 16: Details for Book ID 14915.

PG Metadata	
ID	71451
Title	Handbuch der Pharmakognosie
Author(s)	A. (Alexander) Tschirch, 1856-1939
Language	German
Sources	
Book Content	**1909**. Found in the first pages of the book, with a publication date of 1909.
LoC Catalog	**1909**. One record is found showing 1909 as the publication date.
Google Books	**1909, 2000s**. Multiple records are found, showing 1909 under "Originally published"; some records also show recent publication dates (2000s).
Wikipedia	**1909-1927**. No page is found for the book; the author's page mentions the book and the publication of its four volumes from 1909 to 1927.
Open Library	**1909**. Multiple records are found, most showing 1909 as the publication date.
WorldCat	**1909, 2000s**. Multiple records are found showing 1909 as the publication date; some records show dates from the 2000s.
ChatGPT	**1909**. The source is Open Library and the PG book.
Comments	The book is a popular pharmacognosy text, published in 1909, with four volumes released from 1909 to 1927.

Table 17: Details for Book ID 71451.

B Selected Corpus Statistics

In this section, we show the statistics of our selected corpus. In Figure 5, we plot the distribution of 11 languages in the filtered corpus. In Figure 6, we plot the distribution of books per decade, the numbers of books in the 17th century and the second half of the 20th century are minimal, but they are not zero. In Figure 7, we plot the number of books per century. In Figure 8, we plot the number of tokens per each language in the filtered corpus, this shows that despite the relatively small number of books in some languages, they still contain a significant amount of tokens.

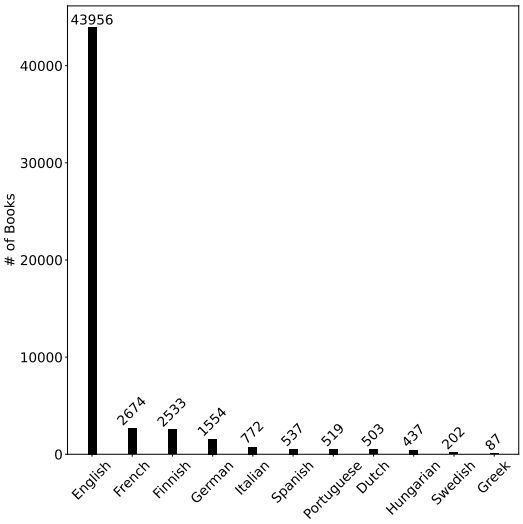

Figure 5: Number of books per each language in the filtered corpus.

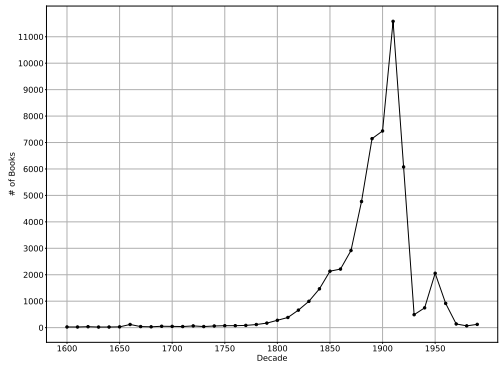

Figure 6: Number of books per each decade in the filtered corpus.

Visualizing the genres distribution in the corpus is tricky as one book can be assigned by multiple genres. In Figure 9, we show the frequency of each genre, as well as the frequency of intersecting genres.

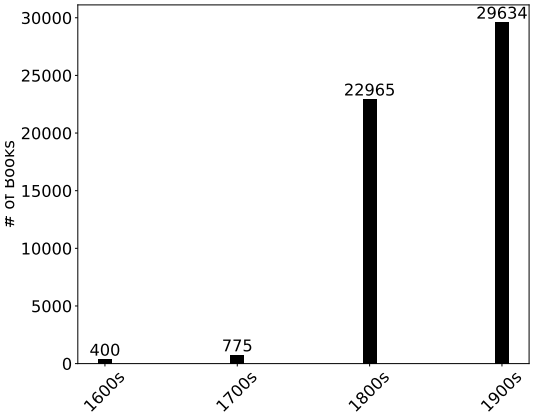

Figure 7: Number of books per each century in the filtered corpus.

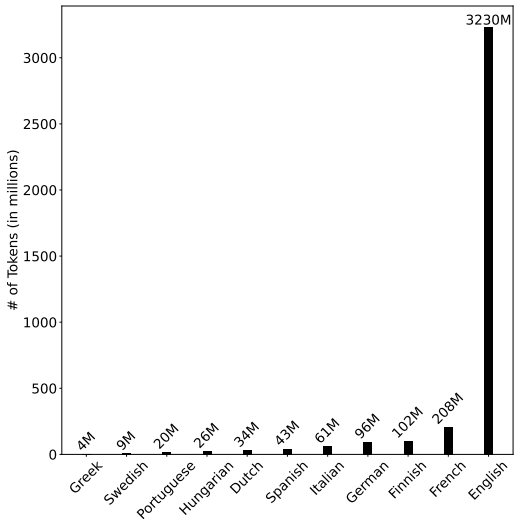

Figure 8: Number of tokens per each language in the filtered corpus.

C Experiments Settings

In this section, we report the settings of all our experiments that are not mentioned in § 5.

C.1 RAG Experiments

For LLM_{web}, we query Searxng with its default search engines selected, using the search query: "the original publication year of <book-title> by <book-author>". The output of the search process is an aggregation of the related snippets found in the top 50 search pages. Then we add the search result to the following prompt before feeding it to the LLM: "You are an accurate AI assistant that can figure out the original year of publishing any book from the Project Gutenberg website. There might be multiple years of publishing for a single book. But we are interested in the oldest year of publishing only. You will be given the title and the author of the book. You will also be given up to date results from the internet related to information about the original publishing year of this book. Please respond with the answer only e.g. '1897'. If you can't find the answer, please respond with 'Not Found'. Title:<book-title> Author:<book-author>, Search Results:<search-results>"

For $\text{RAG}_{\text{wikipedia}}$, we retrieve the top 3 relevant chunks in Wikipedia dumps to the query: "the original publication year of the book titled <book-title> by <book-author>". We then add the retrieved chunks to the following prompt before feeding to a LLM: "You are a great accurate AI assistant that can figure out the original publishing year of any book, you will get the title and author(s) of the book, and additionally the top 3 relevant pieces of text to this book extracted from Wikipedia. The additional extracts from Wikipedia are in a JSON format, it is an output of a smart retrieval system. Based on this information only, you should figure out the year of publishing this book. If you don't know the answer, please respond with 'no answer'. Title: <book-title> Author: <book-author>. Top 3 relevant texts in Wikipeida: <wikipedia-results>

In the RAG_{book} experiment, we chunk each book into chunks of 2 000 characters each, with an overlap size of 100 characters. Then for each book, we query the chunks of the books for the most relevant chunks to the query: "the original publication year of the book titled <book-title> by <book-author>". We then add the most relevant chunks to the following prompt before feeding it to the LLM: "You are a great accurate AI assistant that can figure out the original publishing year of any book, you will get the title and author(s) of the book. Additionally, you will be provided with 3 pieces of text extracted from the book content, which can be useful for you to figure out the year of publishing. Based on this information

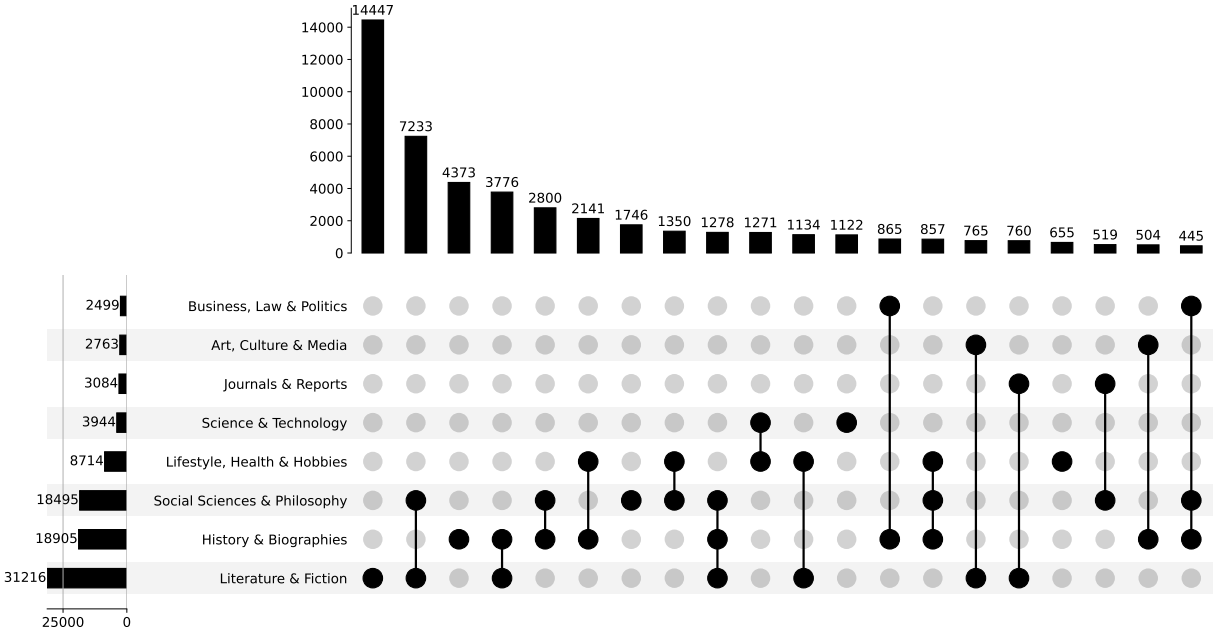

Figure 9: Genres Distribution

only, you should figure out the year of publishing this book. If you don't know the answer, please respond with 'no answer'. Title: <book-title> Author: <book-author>. Top 3 relevant texts in the book content: <book-results>".

C.2 Selected Content Experiments

In Baseline experiments, we used the following prompt: "You are a great accurate AI assistant that can figure out the original date of publishing any book in the Project Gutenberg corpus. Records of Project Gutenberg books only show the release date of the digital version of the book, but we need to get the original date of publishing the first paper version. Please respond with only the year of publishing the book, do not include any other information in your answer. If you don't know the answer, please respond with 'no answer'. What is the original date of publishing the Book: <book-title> by <book-author>?".

In $\text{LLM}_{\text{first}}$, $\text{LLM}_{\text{first/last}}$, and $\text{LLM}_{\text{first/last/summary}}$ we use the following prompts: "You are a great accurate AI assistant that can figure out the original publishing year of any book, you will get the title and author(s) of the book, and extracted text of the book, and based on this information only, you should figure out the year of publishing this book. If you don't know the answer, please respond with 'no answer'. What is the original date of publishing the Book: <book-title> by <book-author>? Extracted Book Content: <selected-content>"

In $\text{LLM}_{\text{whole}}$, we use the following prompt sequentially until we either get an answer or the book ends: "You are an accurate AI assistant that can retrieve the original year of publication of any book from the textbook itself. You will get the title of the book, and in addition to that, you will get an extracted text from the textbook itself. If the original publication year of the book is mentioned in the text, you should respond with the year and the line of text that mentions the year in the provided text. Otherwise, if the publication year is not mentioned in the provided text, please respond with 'no answer'. In your response, you ONLY have two options to respond with: 1. The retrieved year of publication of the book. As well as the sentence that contained that year in the provided text. 2. Or if the year of publishing the book is not mentioned in the provided text, you should respond with: 'no answer' only. You should prioritize accuracy over getting an answer, so if you are not

sure about the year, just respond with 'no answer'. What is the original date of publishing the Book: <book-title> by <book-author>? Extracted Book Content: <selected-content>".

In $\text{LLM}_{\text{detailed-protocol}}$, we use the following prompt: "You are an accurate AI assistant that can figure out the original year of publication of any book. You will get the title of the book, in addition to extracted texts of the first and last pages of the textbook, and based on this information ONLY, you should figure out the original year of publishing this book. Please follow the following protocol in your inference: 1. Search for the publication year in the provided information. 2. Consider also the year if it was interpreted as the year of writing the book by the author. 3. If multiple years are thought to be the publication year, please consider the oldest publication year only. 4. Be aware that years may be written in Roman numerals, so consider Roman numeral years too, but convert them in your response into standard numeric format (e.g., MDCCCLXXX should be interpreted as 1880). 5. Please respond with only the inferred year, do not include any other information or texts. 6. If you can't find any relevant information about the year of publishing or writing the book, please respond with 'no answer' only. 7. NEVER RESPOND WITH A YEAR THAT IS NOT MENTIONED IN THE PROVIDED TEXTS BELOW (FIRST AND LAST PAGES OF THE BOOK). What is the original date of publishing the Book: <book-title> by <book-author>? Extracted Book Content: <selected-content>".

C.3 Open Library API Experiment

For each book in the PG corpus, we query https://openlibrary.org/search.json?<params>, with <params> having the title and the author of the book, and specifying the limit of results to 10. We then look at the field first-publish-year in the response and get the oldest year as the final answer by this method.

D Filters and Heuristics List

The following heuristics and filters are applied to the extend PG corpus to produce a filtered version:

1. Include books where the estimated year appears at least once in the first or last 5,000 tokens of the book content.
2. Include books where the estimated century from the $\text{LLM}_{\text{first/last}}$ experiment matches the century estimated from the automatically generated summary.
3. Exclude books estimated to be written before 1600 or after 2000, as the numbers of published books beyond these periods are very minimal and only causing outliers.
4. Retain only books written in one of the 11 most frequent languages in PG, for the same reason as above.
5. Exclude books with fewer than 1 000 tokens.
6. Retain only books with at least one assigned genre in the PG browsing category metadata.