

Project Gutenberg's Artificial Intelligence Initiatives

Ilana Kingsley

ABSTRACT

This paper reviews Project Gutenberg's (PG) use of artificial intelligence (AI) with the free books in its digital library. The paper covers three main projects where PG has integrated AI into its workflow. The three areas discussed are disparate in nature and include audiobook, book summary, and browsing category generation. The paper reviews the AI models and processes used, as well as how volunteers and users have responded to the organization's use of AI. PG's dataset is extremely large and includes more than 75,000 books. Large datasets can be challenging and costly for AI models to process. The main goal of this paper is to share how PG has incorporated AI processes into the workflow of a very large digital library while maintaining human oversight.

INTRODUCTION

Project Gutenberg (PG) is an online library of more than 75,000 free books that are out of copyright. The Project Gutenberg Literary Archive Foundation (PGLAF) is the 501(c)3 nonprofit charitable organization that runs Project Gutenberg. PGLAF is run mostly by volunteers and receives donations to support its cause. The organization has always been at the forefront of embracing technology and continues to do so by exploring generative artificial intelligence (AI). At present, PG has three main AI initiatives: (1) an open audiobook collection that uses AI to narrate books with cadence and emotion, (2) AI-generated book summaries, and (3) AI-generated browsing categories. These initiatives are disparate in nature. The audiobook collection uses technology developed by Microsoft and the Massachusetts Institute of Technology (MIT), as well as prompt engineering, and its intent is to improve AI text-to-speech (TTS) applications so that text is spoken with cadence and emotion. The AI-generated book summary initiative is similar to the browsing category initiative, as they both use prompt engineering to generate text-based summaries or categories. A book summary is defined as a description or synopsis of a book, whereas a browsing category classifies the overall subject of a book. Prompt engineering is a way of guiding large language models (LLMs) or generative AI models to output a desired result.

A key factor of PG's AI initiatives is to maintain human oversight and to invest the time needed to continually revise prompts in order to achieve high quality. PG places a strong emphasis on maintaining high standards and is cautious about using AI unless the quality of audiobooks, book summaries, and browsing categories is reliably and consistently excellent.

This paper is divided into three main sections: audiobooks, book summaries, and browsing categories. Each section focuses on PG's use of generative AI.

AUDIOBOOKS

Traditionally, audiobooks were created for accessibility reasons to help persons with low or no vision.¹ By the mid-2020s, audiobooks are increasing in popularity as a way for people to consume books.²

About the Author

Ilana Kingsley (imkingsley@alaska.edu) is Professor of Library Science and Web Librarian, University of Alaska Fairbanks. © 2026.

Submitted: 1 May 2026. Accepted for Publication: 5 June 2026. Published: 21 September 2026.

The ability to listen to PG books is not new. PG books are published in text-based formats, including plain-text and HTML. Screen readers, such as JAWS or NVDA for Windows and VoiceOver for macOS, are able to read books aloud; however, the sound is robotic. PG has a number of computer-generated audiobooks from 2003, but as noted on PG's homepage, "They are computer-generated and listenable but relatively low quality compared to today's technology."³ The PG library also contains 622 audiobooks read by human volunteers from LibriVox. LibriVox is a nonprofit, crowdsourced organization where humans volunteer their time to read and record books that are available in the public domain.⁴ Each PG LibriVox book is read by a human. Unlike the robotic sounds of screen readers, because LibriVox audiobooks are human-read, they have excellent prosody, meaning the narrator's voice has natural rhythm, tone, intonation, pitch, and tempo.

The use of AI to create audiobooks from PG books began in 2023.⁵ In her 2025 conference presentation, Kingsley provides three examples of PG audiobooks. Each example plays a snippet from Jack London's *The Call of the Wild*. In the first example, audio is narrated by a human.⁶ The second example, from 2004, is a computer-generated narration.⁷ The third example, from 2023, is the AI Microsoft/MIT/PG model.⁸ In 2026, Microsoft/MIT/PG updated their TTS/AI models, and *The Call of the Wild* is now available in female or male synthetic voices.⁹ It's been noted that the 2023 AI audiobooks are an improvement from computer-generated narration, but "there is still a nonhuman element to the audio voice, one that lacks inflection."¹⁰ With regard to the 2026 audiobooks, Wong determined that human voices ranked higher than AI voices, and that listeners preferred the newer 2026 synthetic voices over the 2023 synthetic voices.¹¹

Audiobook Literature Review

PG's collection is mostly books. Greg Newby, the late CEO of PG, understood that there was a need for audiobooks, but that creating audiobooks was beyond the capacity of the small volunteer organization.¹² That's when Newby partnered with a team from Microsoft to use AI neural TTS technology to generate audiobooks from the PG collection.¹³

A resulting conference paper, "Large-Scale Automatic Audiobook Creation" describes the technology and process used to generate audiobooks from the PG library in 2023. In this model, HTML-based ebooks are converted to audiobooks using SynapseML.¹⁴ SynapseML is Microsoft's "open-source library that simplifies the creation of massively scalable machine learning (ML) pipelines."¹⁵ In simplified terms, to generate audiobooks, ebooks are loaded into Microsoft's Azure Synapse Analytics platform, parsed using SynapseML, and then audio is created using Azure's AI neural TTS services.¹⁶ Since the first release of 5,000 audiobooks in 2023, the team has been focusing on improving the quality of AI speech synthesis. They are working to enhance the naturalness, clarity, and emotive expressiveness of synthetic voices. Wong's 2026 prepublication draft conference paper discusses an updated model for generating audiobooks using the LLM GPT-4.1 for text processing and gpt-4o-mini-tts for converting processed text into spoken audio.¹⁷

Besides the PG audiobook initiative, other researchers are working to improve TTS AI applications. Kryvenchuk and Duda provide an overview of current TTS synthesis and audiobook creation methods using AI and machine learning. One of the focuses of their article discusses using OpenAI GPT-4 to segment narration into three parts: that of the author's voice (the narrator and commentator), male voices for male characters, and female voices for female characters.¹⁸ Magadán-Díaz provides a good outline of what it takes to produce an audiobook from text using AI, which includes preprocessing text, configuring an appropriate synthetic voice for narration,

and adjusting prosody.¹⁹ Her article also reports on a study that identifies proprietary generative AI tools used to create synthetic voices in the audiobook publishing industry.²⁰

With regard to listener preference, recent studies have found that human-narrated audiobooks received slightly higher ratings than AI-narrated audiobooks; however, listeners do enjoy AI-narrated audiobooks.²¹

Audiobook Challenges and Discussion

The challenges of generating audiobooks using AI are varied. They include, but are not limited to, issues such as large-scale production, cost, ethics, paratext, and listener preference. These challenges are discussed next.

Large-scale Production

PG has a very large and free dataset, making it appealing for researchers from all over the world to test new TTS and AI applications. Processing large amounts of data can be challenging for many reasons. With regard to LLMs, challenges include limited tokens and context length, computational and memory constraints, and performance degradation, including hallucinations.²² Other challenges revolve around the structure of the text, such as sentence length, style, and type; dialogue; and speaker style.²³ Using prompt engineering, Wong and colleagues have successfully processed more than 30,000 PG books into audiobooks and have demonstrated that large-scale production of high-quality audiobooks is possible.²⁴

Cost and Ethics

Traditional methods of audiobook creation are time-consuming and costly, as they include hiring professional narrators.²⁵ The use of AI for audiobook generation eliminates the need for human narrators and drastically reduces production costs. However, eliminating human narrators contributes to the ethical implications of job loss, voice cloning issues, and other forms of human exploitation.²⁶

With regard to PG audiobook creation, Newby noted that its volunteer-driven team did not have the capacity to produce high-quality audiobooks.²⁷ In a conversation with him, Newby said that the benefits of using AI to produce audiobooks outweighed the drawbacks—that there was a user-based need to produce audiobooks, and without having a volunteer base to narrate and create such books, AI could help fill the gap.

Paratext

PG books contain paratext. Paratext is defined as text that relates to the main body of a book and provides readers with additional information about the book.²⁸ PG books contain front and back paratext.²⁹ They may also have figures, charts, and tables within the main body. Front paratext includes information about the use of the book, the release date, credits, and more. Back paratext includes legal information. Paratext does not need to be narrated as it's not an essential part of the book. Paratext is a challenge for AI audiobook creation.³⁰ Walsh's system handles paratext by analyzing and parsing HTML files.³¹ Wong's approach uses LLM prompt engineering to classify text into different parts, such as chapter headers, readable text, or paratext.³²

Listener Preference and Usage

Some studies have shown that listeners prefer human-narrated audiobooks over synthetic-narrated voices.³³ With the rapid improvement of AI voice synthesis, this preference is narrowing.³⁴

Regarding the 2023 set of PG AI-narrated audiobooks, there was not much listener feedback, except that users noted that the books were all narrated in a male voice.³⁵ Walsh's audiobook demonstration at Interspeech 2023 had the ability to personalize voices using voice cloning.³⁶ Consequently, the demo was able to narrate books in male and female voices, but this feature was not integrated with the 5,000 PG audiobooks. The newer 2026 release of PG audiobooks, available at torrentialstorm.com/audiobooks/library, has incorporated female and male AI narrators, albeit using the same narrator's voice for each book.

PG AI-narrated audiobooks reside on non-PG servers. The 5,000 audiobooks released in 2023 reside on The Project Gutenberg Open Audiobook Collection website, which is owned by Mark Hamilton from MIT and Microsoft. The 2006 collection of audiobooks resides on torrentialstorm.com, which is owned by Gandalf Hernandez and is a temporary location; eventually the audio files will be transferred to a PG server. Because the AI audiobooks currently reside on non-PG servers, it's difficult to access user statistics. Results from the feedback form on The Project Gutenberg Open Audiobook Collection website are currently sent to Microsoft's Brendan Walsh and not to PG. At the project's inception, feedback was shared with Newby, who reviewed audio quality.

Audiobook Conclusion

The publishing industry has seen an increase in popularity in audiobooks.³⁷ With this increase, PG recognized the need to produce audiobooks but did not have the volunteer capacity to create them. As technology advanced and LLMs were able to process large-scale, text-based datasets, and with the help of MIT and Microsoft, generating audiobooks from the PG library became a reality. There are more than 35,000 PG-based audiobooks in MP3 format available to download for free. Future directions of PG audiobooks include processing the entire PG corpus and developing multi-speaker audio narration.³⁸

BOOK SUMMARIES

For PG, a book summary is a brief synopsis or description of a book. PG book summaries are similar to those of traditional book descriptions, sometimes known as back-cover copy or flap copy. There are nearly 75,000 PG AI-generated book summaries that appear at the top of the page for each book. Below is an example of an AI-generated book summary from *The Call of the Wild* by Jack London. At the end of the summary, it's noted that "this is an automatically generated summary."

"The Call of the Wild" by Jack London is an adventure novel published in 1903. Buck, a powerful dog living comfortably in California, is stolen and sold into service as a sled dog during the Klondike Gold Rush. Thrust into the brutal Yukon wilderness, he must fight to survive among other dogs and harsh masters. As Buck endures the unforgiving environment, he grows increasingly primitive, shedding his domesticated life and answering an ancient instinct that pulls him toward something wild and primal. (This is an automatically generated summary.)

Human-written summaries take time to write. PG volunteers who digitize books are not interested in writing book summaries, as they would prefer to spend their time digitizing books so they can expand the PG corpus.³⁹

PG introduced AI-generated book summaries in 2024, as CEO Newby and the PG board thought summaries would enhance the collection by helping readers decide whether or not the book was the right fit for them.

Book Summaries Literature Review

Traditionally, back-cover copy, also known as dust jacket copy, was an important part of advertising and potentially one of the only ways to draw a reader to purchase a book.⁴⁰ Some research has been conducted on predicting reader satisfaction using book descriptions in online bookstores.⁴¹ There has been much research in the area of automatic text summarization (ATS), prompt engineering, and LLMs, but not directly in relation to generating short book summaries as defined by PG. Research about book summarization indicates that ATS systems, which are developed to summarize short documents, do not scale for books.⁴²

With regard to prompt engineering and LLMs, few-shot instruction-based prompt engineering is the method used to generate PG book summaries. Few-shot prompts include a limited number of examples, which help the AI model with comprehension and accuracy.⁴³ Such prompts also help the model learn and understand patterns it is expected to complete.⁴⁴ Compared with one-shot and zero-shot techniques, few-shot prompt engineering has been proven to have a higher success rate in text summarization.⁴⁵ Techniques used to craft efficient and effective prompts include, but aren't limited to, avoiding ambiguity, being concise and specific, providing contextual information and examples, specifying output format, and continuously iterating and refining.⁴⁶

Book Summaries Challenges, Prompts, and Feedback

This section discusses challenges, prompts, user feedback, and human oversight with regard to PG book summaries.

Challenges and Prompts

PG's programmer, Johannes Seikowsky, encountered only a few challenges generating book summaries using prompt engineering. The main challenge was cost. In an email exchange with Seikowsky, he said that cost was a prohibitive factor because PG has such a large dataset and a slim budget. ChatGPT 4.0 mini was the LLM originally used in 2024 and was not free. PG's use case required a long section to input data as well as a long prompt, which called for a paid subscription. The cost of the ChatGPT subscription was approximately \$1,000 USD.⁴⁷ Besides cost, context limitations were an issue. When the original set of summaries was generated, the LLM was unable to process book-length texts due to their size, as well as the limited size of the LLM's context window. With recent advances in LLMs, context limitations are still present but becoming less of an issue.

To generate accurate book summaries, crafting prompts is extremely important and takes multiple iterations and patience. Due to context size limitations, Seikowsky's prompt instructed the AI model to look at only the opening portion of a book, approximately the first 12,000 words. This meant that for some short stories the AI model examined the entire text, whereas for lengthy books it did not. Based on the AI model's understanding of the input, it was then instructed to write two paragraphs that were concise and to the point. The first sentence told the AI model to include the title of the book, the name of the author, the type of book (e.g., a novel, scientific publication, collection of short stories, etc.), and the time period in which the book was written. The prompt offered examples of what the first paragraph should look like. After the system generated the first paragraph, it was told to write a second paragraph based on the first 12,000

words and was given instructions on how to do this. An example prompt used to create book summaries is in Appendix 1.

User Feedback and Prompt Revision

Much of the user feedback received about book summaries has been positive, although some of it has been negative. Feedback is emailed to PG using one of the email addresses listed on its Contact Us webpage. Some of the feedback received was about problems with content and parsing errors, while other feedback was about the general use of AI.

Content errors include incorrect summaries, parsing errors, and title/summary mismatches.⁴⁸ *The Flirt* by Booth Tarkington and *Amours De Voyage* by Arthur Hugh Clough are examples of two books that had incorrect summaries. The AI stated they were written in the 19th century, when in fact they were written in the 20th century. An example of a parsing error was when an error message was displayed instead of a book summary. In this case, the parser didn't work with RTF-format files; it only worked with HTML and plain text. An example of a title/summary mismatch was when the AI output a title as *Tutankhamen and the Discovery of His Tomb*, but the summary was about a children's adventure story.

Summaries are corrected when complaints are lodged. The correction is to update the prompt and rerun the script for that specific book. The updated prompt is used in the future for summarizing newly added books. Existing summaries are not updated unless there's a reason to do so.

With regard to using AI, some users have provided positive feedback such as that the LLM-generated summaries are a nice addition to the site. Others have shared negative feedback such as complaints of summaries being distracting, verbose, and unnecessary.

In an email exchange with Newby's successor, CEO Eric Hellman said that a recent improvement to book summaries, based on user feedback, is that they are no longer generated using the Appendix 1 prompt if there's an existing Wikipedia article about the book. This is because summaries are more refined and accurate when the LLM summarizes a book using its Wikipedia article. For books that have existing Wikipedia articles, book summaries are generated using a shorter prompt (see Appendix 2). The LLM used to accomplish this is Sonnet-4.5.

Human Oversight

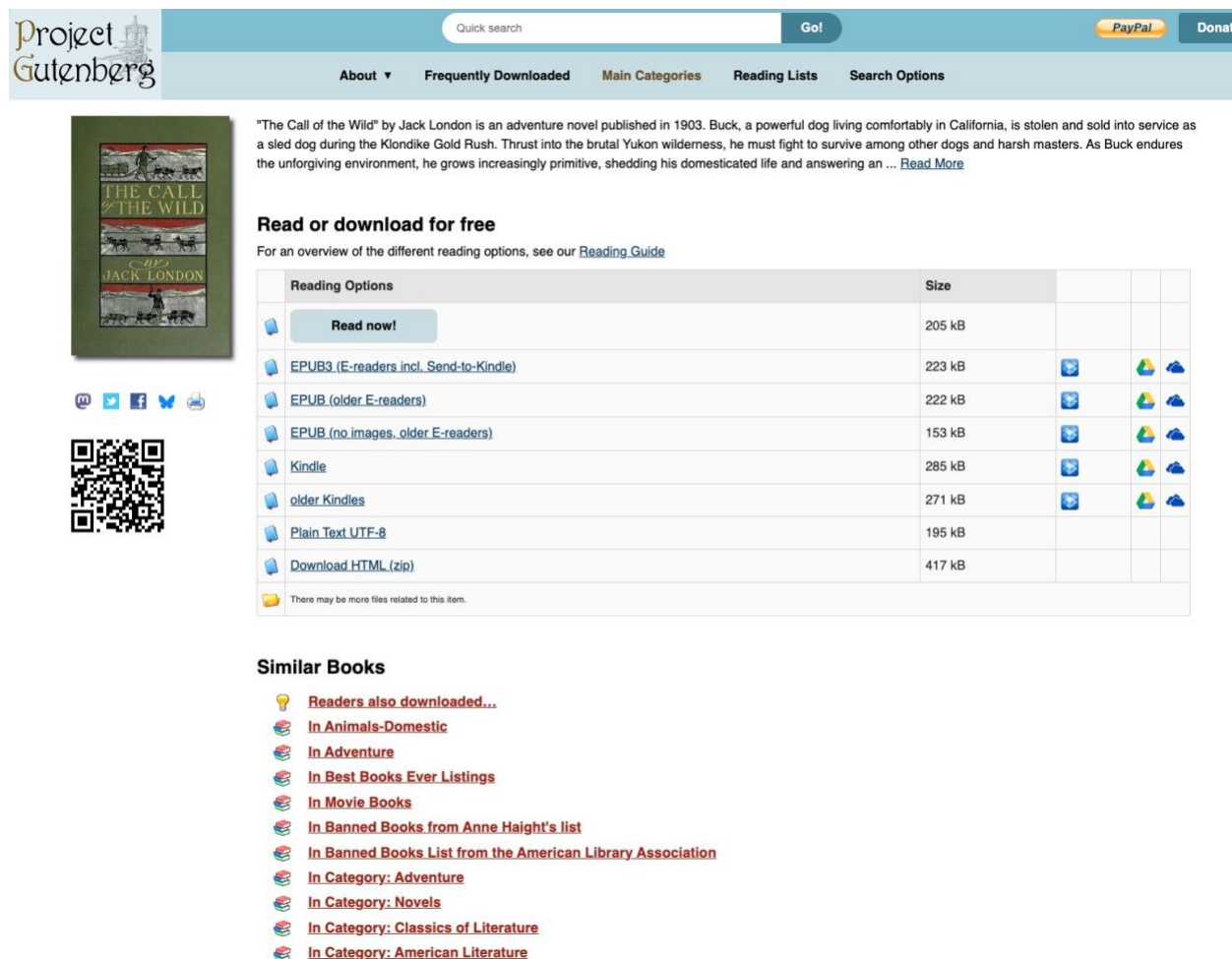
When the book summary prompt was originally launched, Newby spent time reviewing output for quality control, and PG's programmer Seikowsky spent a considerable amount of time revising the prompt. When they were confident that the model was generating accurate summaries, they eased off reviewing the LLM's output and began relying on a crowdsourced approach to report content errors and anomalies. PG users let the team know when a summary is incorrect and needs to be revised, as well as whether or not they are fans of AI.

BROWSING CATEGORIES

Browsing categories are commonplace in physical and online bookstores. They are a way of organizing books so that customers can quickly and easily find them. Unlike traditional library subject headings, which usually follow a detailed hierarchical controlled vocabulary, browsing categories are flatter and provide a high-level descriptor of what a specific book is about. On the PG website, browsing categories are labelled Main Categories and are similar to those you would find in a bookstore. The initial list of categories found on the page is human-generated. For example, on the Main Categories page, available at <https://www.gutenberg.org/ebooks/categories>, the Literature category and its subcategories

were created by a human. AI is used to assign categories to each book in the collection. For example, *The Call of the Wild* has ten AI-assigned categories listed under Similar Books, as depicted in Figure 1.

Figure 1. Example AI categories assigned to *The Call of the Wild*.



The screenshot shows the Project Gutenberg website interface. At the top, there is a navigation bar with links for 'About', 'Frequently Downloaded', 'Main Categories', 'Reading Lists', and 'Search Options'. A search bar is also present. The main content area features the book cover for 'The Call of the Wild' by Jack London, a QR code, and a brief description of the book. Below the description, there is a section titled 'Read or download for free' with a link to the 'Reading Guide'. A table lists various reading options and their sizes. At the bottom, there is a 'Similar Books' section with a list of categories.

Reading Options	Size
Read now!	205 kB
EPUB3 (E-readers incl. Send-to-Kindle)	223 kB
EPUB (older E-readers)	222 kB
EPUB (no images, older E-readers)	153 kB
Kindle	285 kB
older Kindles	271 kB
Plain Text UTF-8	195 kB
Download HTML (zip)	417 kB

Similar Books

- [Readers also downloaded...](#)
- [In Animals-Domestic](#)
- [In Adventure](#)
- [In Best Books Ever Listings](#)
- [In Movie Books](#)
- [In Banned Books from Anne Haight's list](#)
- [In Banned Books List from the American Library Association](#)
- [In Category: Adventure](#)
- [In Category: Novels](#)
- [In Category: Classics of Literature](#)
- [In Category: American Literature](#)

The main reasons for using AI to categorize books are the lack of volunteer capacity and speed of categorization. Using AI for bulk classification is much faster than what humans can accomplish. PG volunteers used to categorize books, but at a rate of 200 books per month, it is too much work for volunteers.

It's important to note that PG has a dedicated volunteer cataloger who assigns Library of Congress Subject Headings (LCSH) to each book. As the sole cataloger, they have a high workload and are not interested in the additional task of assigning high-level browsing categories. Another element related to PG's browsing categories is human-curated reading lists, available at <https://www.gutenberg.org/ebooks/bookshelf>.

Browsing Categories Literature Review

The use of technology to classify books is not new. Many institutions are experimenting with automatic subject indexing, but libraries have not yet adopted full-scale use of computer-assigned subjects and opt for machine-aided or computer-assisted indexing in which humans are involved in the decision-making process.⁴⁹ Since the emergence of AI, libraries and similar organizations

are using AI to help with the classification of books and documents. Applications such as CatalogerGPT have been developed to generate MARC records and assist with assigning LCSH tags.⁵⁰ BERT (Bidirectional Encoder Representation from Transformers) has been successfully used as a machine-aided subject indexing tool to classify PG's books with LCSH.⁵¹ AI has been used for thematic classification and assignment of the United Nations' Sustainable Development Goals to open access documents and has proven to be a useful tool for assisting librarians, but not as a replacement for them.⁵² These are just a few examples of the many projects that are experimenting with using AI to classify books and documents.

Browsing Categories Challenges and Prompts

This section discusses challenges, prompts, user feedback, and human oversight in regard to PG book categories.

Challenges and Prompts

Prompt engineering is time-consuming, takes patience, and requires much iteration. Seikowsky said that it's important to tell the model the purpose of the outcome. For example, the purpose of browsing categories is for the categories to be displayed on a website. It's also necessary to inform the AI of common mistakes; for example, if a book is assigned to "History – Warfare" (or any other history category), it should not be assigned to "History – Other" as well.

Similar to creating PG book summaries, a few-shot instruction-based prompt is used to generate browsing categories. Unlike the book summaries prompt, the browsing categories prompt does not scan the first 12,000 words of a book. Instead, it uses the book summary to classify the book. Hence, it's important that the book summary is accurate. An example prompt used to generate browsing categories is in Appendix 3.

User Feedback

There has been no user feedback regarding browsing categories.

Human Oversight

Similar to the book summary prompt, when the browsing categories prompt was launched, Newby spot-checked category assignment and Seikowsky updated the prompt as needed.

CONCLUSIONS

PG has integrated AI into its workflow in the form of generating audiobooks, book summaries, and browsing categories. Based on positive user feedback, and a relative lack of negative feedback, these three disparate projects have been successful. Reasons for incorporating AI into the PG workflow include volunteer capacity limits and the need to evolve with emerging technological paradigm shifts. PG views the use of AI as a tool to enhance the accessibility of its collection and understands that human oversight for quality control is essential. The team will continue to use AI and experiment with new technologies.

ACKNOWLEDGMENTS

This paper is written in memory of Gregory B. Newby, Project Gutenberg's first CEO.

Thank you to Johannes Seikowsky for sharing his prompts.

APPENDIX 1: EXAMPLE PROMPT USED TO CREATE PG BOOK SUMMARIES

```
def summarise_beginning_of_book(title_and_author, text):
    system_prompt = {"role": "system", "content": """"You are a helpful assistant that is very
    good at deriving an understanding of books based on their opening chapters. You are also
    very good at writing texts about those books based on your understanding that are useful
    to potential readers who need to decide whether a particular book is interesting to them
    or not."""}

    prompt1 = {"role": "user", "content": f""""You will be given the opening portion of a
    book. Read it very carefully to understand its content and derive an idea of the book in
    general. Based on your understanding write two paragraphs.
```

The first paragraph should be concise and to-the-point. The first sentence should include:

- the title of the book and name of the author (if provided), which is: {title_and_author}.
- what type of book it is (e.g. novel, scientific publication, historical account, collection of short stories etc.)
- what general time period the book was probably written in (e.g. early 19th century, mid-19th century, late 19th century, 14th century). Only refer to general time periods like that, do NOT use specific years and dates!

Overall the general structure of the first sentence should be exactly this: "<title_name>" by <author_name> is a <book_type> written in <time_period>.

However only include these values if you're reasonably sure. If there is no author_name or you can't reasonably tell the time_period for example, just don't include them.

After that first sentence finish the first paragraph with a concise statement about what the likely topic of the book is. If it's a fiction work mention the main character(s) if you can (but avoid spoiler alerts). The goal of this is that the reader gets a first impression of the book that he or she can use to decide whether the book seems interesting or not.

As said, this first paragraph should be concise. 4 sentences or less is best.

Here are some EXAMPLES of how this first paragraph should principally look like:

- 1) "Adventures of Sherlock Holmes" by Sir Arthur Conan Doyle is a collection of detective stories written during the late 19th century. The book follows the brilliant detective Sherlock Holmes and his companion Dr. John Watson as they solve various cases.
- 2) "The Elements of Style" by William Strunk Jr. and E.B. White is a guidebook on English language usage written in the mid-20th century. The book lays out the principles of clear and concise writing, offering rules and recommendations for effective communication.
- 3) "The Art of War" by Sun Tzu is a treatise on military strategy written in the 5th century BC. The book lays out principles of military strategy and tactics, emphasizing the importance of intelligence, deception, and the use of terrain.
- 4) "Introduction to Algorithms" by Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein is a comprehensive computer science textbook written in the late 20th century. The book follows a structured approach to teaching algorithms, covering a wide range of topics from basic data structures to advanced algorithmic techniques.
- 5) "Vesper Talks to Girls" by Laura A. Knott is a collection of motivational addresses written in the early 20th century. The book compiles talks given to young women at Bradford Academy, addressing aspects of personal development, friendships, and the importance of character.

After you finished this first short paragraph, please write a second paragraph with a concise summary of the book's opening portion that I will provide you with.

In that second paragraph focus on the actual content of that opening part (i.e. the characters, the content, the storyline etc.).

Make sure to cover only the main points and main ideas, whilst avoiding any unnecessary information or repetition.

Overall this summary should be quite brief.

Make it very clear that this is a summary of only the opening portion of the book. Do that by starting the paragraph with a variation of "The beginning of ..." or "The opening of ..." or "At the start of ...".

In your writing use simple, clear, concise and expressive language.

State the name of the book and the author's name only once throughout your response.

In some cases there may be a more appropriate word than "book" to refer to the work I'll give you. Use your judgement to use appropriate terminology. Always write in English."""}

```
prompt2 = {"role": "assistant", "content": "Understood! Please provide the opening portion of the book and I will follow your instructions."}
```

```
prompt3 = {"role": "user", "content": "START OF BOOK BEGINNING: \n" + text + "\nEND OF BOOK BEGINNING"}
```

```
response = openai_client.chat.completions.create(model="gpt-4.1",  
messages=[system_prompt, prompt1, prompt2, prompt3])  
response = response.choices[0].message.content  
return response
```

Credit to Johannes Seikowsky.

APPENDIX 2: EXAMPLE PROMPT USED TO CREATE PG SUMMARIES FROM A WIKIPEDIA ARTICLE

SYSTEM_PROMPT = "You are skilled at writing compelling, teaser-style introductions for literary and artistic works that intrigue readers without revealing everything."

USER_PROMPT_TEMPLATE = ""The official title of this work on Project Gutenberg is: {gutenberg_title}"

Please use this as the authoritative title in your introduction, even if the Wikipedia article uses a different title (this may occur due to language differences or variations in naming).

I'm going to give you the first 1200 words of the Wikipedia article of this work. Write an introduction-text for it that works like a movie trailer – giving a sense of what the work is about without revealing too much i.e. avoid spoilers. Write between 80-90 words. Do not exceed 90 words.

If possible, the first sentence should follow this pattern: "(title)" by (author/composer) is a (type of work) written/published/composed in (time period).

Note: The time period should be as specific as possible. If you have a specific year use that. If you have a decade or century, use that. If you don't have relevant information, please omit it.

Base your word choice of using "written", "published", or "composed" on the content of the Wikipedia article.

Examples:

- "To Kill a Mockingbird" by Harper Lee is a novel written in 1960.
- "Symphony No. 9" by Ludwig van Beethoven is a symphony composed in 1824.
- "Letters to a Young Poet" by Rainer Maria Rilke is a collection of letters written between 1902-1908.

Keep your writing very simple and straightforward. Avoid long sentences and unnecessarily complex language. Be clear and direct.

Make sure to base EVERYTHING on the content of the Wikipedia article! Do NOT add any information that is not presented in the Wikipedia article.

Always write in English, even if the Wikipedia article is in a different language.

It is possible that the Wikipedia article provided does not give you enough information to write a reasonable introduction-text based solely on it. It may be very short or may actually be about the author rather than the work or have some other flaw. At any rate, if the article doesn't provide you with enough information about the work itself, do NOT write an introduction text, but instead just return "insufficient information".

```
<WIKIPEDIA ARTICLE>
{article_text}
</WIKIPEDIA ARTICLE>""
```

Credit to Johannes Seikowsky.

APPENDIX 3: EXAMPLE PROMPT USED TO CREATE PG BROWSING CATEGORIES

```
def get_categories(book_id, summary):
    system_prompt = f"""You are an expert at choosing appropriate categories for books. You do
    this by thoroughly read the summary of a book to cultivate an understanding of what the
    book is about. In doing so you think critically about what the central topic or topics of
    the book are and what the peripheral topic or topics are. Just because certain "matching
    words" appear in the summary does not mean that the book is about that topic. You must
    read the summary carefully and think about what the book is actually about. Then and only
    then you pick appropriate categories for the book from this particular list of available
    categories:

    <categories>
    {categories}
    </categories>
    """

    prompt = f"""I will give you the summary of a book. Your job is to read that summary
    thoroughly and mull it over. Then, based on your understanding, use your best judgement
    to pick the most appropriate categories from the list of categories.

    The purpose of this is that the book will be displayed on a website that has these
    particular categories. Bear that in mind when picking the categories. So if you assign a
    category to a book it should be in such a way that if a website user were to browse that
    category on the website, he/she should NOT be surprised to find that book in that
    category. So it's important that you assign each book to the categories that make the
    most sense in light of what the book is about.

    Think sceptically and carefully about what categories are actually appropriate. For
    example, if a certain book is about "animal carvings" that book may appropriately
    assigned to "Art" but not to "Nature/Gardening/Animals". Also be careful to distinguish
    between "Literature" works and "History" works. By "Literature" we in general mean
    fictional writing. Whilst "History" is non-fictional writing about historical
    events/occurrences.

    The category "How To ..." is for books that centrally give practical instruction or
    guidance for a certain task/endeavour.

    The "Classics of Literature" category is for the famous and significant works like "Moby
    Dick", "The Catcher in the Rye", "The Illiad" etc.

    The "Biographies" category is for all works that are generally biographical in nature, not
    only those that are written specifically as "pure" biographies. Diaries, personal
    journals etc. should therefore be included.

    The "Journals" category is for magazines and publications, not personal journals.

    If a book can reasonably be assigned into the categories of "American Literature", "British
    Literature", "German Literature", "French Literature" or "Russian Literature" in addition
    to what type of literature it is, then it should be assigned to both.

    The categories "Literature - Other" and "History - Other" are exclusively for history and
    literature books that do NOT fit into any of the other literature or history categories
    respectively.

    For example, if a book is assigned to "History - Warfare" (or any other history category),
    do NOT assign it to "History - Other" as well. Another example, if a book is assigned to
    "Poetry" (or any other literature category), do NOT assign it to "Literature - Other" as
    well.
    """

    assistant_reply = "Ok, please give me the summary of the book. I will read it thoroughly
    and based on my understanding pick the categories that are relevant for this book. When
    picking the categories I will be bear in mind that these will be used to by people to
    find books that are relevant to them. So I'll make sure to pick categories according to
    what people may be expect to find in each category."
```

```
schema = {
    "name": "book_categories",
    "schema": {
        "type": "object",
        "properties": {
            "categories": {
                "type": "array",
                "description": "Chosen category or categories for the particular book.",
                "items": {
                    "type": "string",
                    "enum": categories_list
                }
            }
        },
        "required": [
            "categories"
        ],
        "additionalProperties": False
    },
    "strict": True
}

# title_and_author = fetch_title_and_author_from_website(book_id)
book_info = f"<summary> {summary} </summary>"

completion = openai_client.beta.chat.completions.parse(
    model="gpt-4.1",
    messages=[
        {"role": "system", "content": system_prompt},
        {"role": "user", "content": prompt},
        {"role": "assistant", "content": assistant_reply},
        {"role": "user", "content": book_info}
    ],
    response_format={"type": "json_schema", "json_schema": schema}
)
category_choice = json.loads(completion.choices[0].message.content)["categories"]
return category_choice
```

Credit to Johannes Seikowsky.

ENDNOTES

- ¹ Matthew Rubery, "Audiobooks, Literature, and Sound Studies," in *Audiobooks, Literature, and Sound Studies* (London, UK: Taylor & Francis Group, 2011), 5.
- ² Michelle Faverio and Andrew Perrin, "Three-in-Ten Americans Now Read e-Books," Pew Research Center, January 6, 2022, <https://www.pewresearch.org/short-reads/2022/01/06/three-in-ten-americans-now-read-e-books>; Audio Publishers Association, "The Audio Publishers Association Annual Survey Shows Audiobook Sales Grow to 2.22 Billion Dollars," Research Surveys Press Releases, accessed April 6, 2026, <https://www.audiopub.org/surveys>; "Innovation Fuels Thriving Audio Industry," *Publisher's Weekly*, June 2, 2023, 35; Marta Magadán-Díaz et al., "Audiobooks and Artificial Intelligence: Tools for Synthetic Audiobook Creation and Implications for the Publishing Industry," *Publishing Research Quarterly* 41, no. 3 (2025): 406, <https://doi.org/10.1007/s12109-025-10049-1>.
- ³ Project Gutenberg, "Project Gutenberg," accessed April 5, 2026, <https://www.gutenberg.org>.
- ⁴ LibriVox, "About LibriVox," accessed April 5, 2026, <https://librivox.org/pages/about-librivox>; Millicent Weber, "'Reading' the Public Domain: Narrating and Listening to Librivox Audiobooks," *Book History* 24, no. 1 (2021): 210, <https://doi.org/10.1353/bh.2021.0007>.
- ⁵ Ilana Kingsley, "Innovating with AI: Project Gutenberg's Approach to Free eBooks and Literary Management," virtual conference presentation, August 18, 2025, by Pacific Northwest Library Association, YouTube, 18:38, <https://www.youtube.com/watch?v=WXD7rUMyzc4>; Microsoft, "Project Gutenberg Brings Free Audiobooks to the Global Community with AI-Generated Book Narration," July 18, 2023, <https://www.microsoft.com/en/customers/story/1646266241611394912-project-gutenberg-nonprofit-azure-synapse-analytics-azure-ai-services>. Eric Oliver, "Why So Many Americans Believe in So Many 'Crazy' Things," moderated by Andrew McCall, virtual lecture, February 23, 2022, posted March 21, 2022, by University of Chicago, YouTube, 1:01:45, <https://youtu.be/hfq7AnCF5bg>.
- ⁶ Kingsley, "Innovating with AI," 4:37.
- ⁷ Kingsley, "Innovating with AI," 4:55.
- ⁸ Kingsley, "Innovating with AI," 5:15.
- ⁹ Female AI narration, *The Call of the Wild* by Jack London, Torrential Storm, accessed April 5, 2026, <https://torrentialstorm.com/audiobooks/audiobook/?id=215-shimmer>; Male AI narration, *The Call of the Wild* by Jack London, Torrential Storm, accessed April 5, 2026, <https://torrentialstorm.com/audiobooks/audiobook/?id=215-ash>.
- ¹⁰ Todd Miller, "Weighing the Merits of AI for Information Access," *Computers in Libraries* 44, no. 2 (2024): 14.
- ¹¹ Stella Wong et al., "Large Scale Audiobook Synthesis Using LLMs and Text-to-Speech," draft conference paper submitted to Interspeech 2026, 4.
- ¹² Microsoft, "Project Gutenberg Brings Free Audiobooks to the Global Community."
- ¹³ Microsoft, "Project Gutenberg Brings Free Audiobooks to the Global Community."
- ¹⁴ Brendan Walsh et al., "Large-Scale Automatic Audiobook Creation," paper presented at Interspeech 2023, Dublin, Ireland, August 20–24, 2023, 3675, https://www.isca-archive.org/interspeech_2023/walsh23_interspeech.pdf.
- ¹⁵ Microsoft, "SynapseML," accessed April 5, 2026, <https://microsoft.github.io/SynapseML/docs/Overview>.
- ¹⁶ Microsoft, "Project Gutenberg Brings Free Audiobooks to the Global Community."
- ¹⁷ Wong et al., "Large Scale Audiobook Synthesis," 1.
- ¹⁸ Yurii Kryvenchuk and Oleksandr Duda, "Audio Book Creation System Using Artificial Intelligence," paper presented at the 2024 IEEE 19th International Conference on Computer Science and Information Technologies, Lviv, Ukraine, October 16–19, 2024, 2, <https://doi.org/10.1109/CSIT65290.2024.10982688>.

-
- ¹⁹ Magadán-Díaz et al., "Audiobooks and Artificial Intelligence," 408–10.
- ²⁰ Magadán-Díaz et al., "Audiobooks and Artificial Intelligence," 411–19.
- ²¹ Annika Ahrens-Schwabe et al., "Who Cares about Artificial Intelligence? Human and Artificial Voices in Audiobooks," *Computers in Human Behavior Reports* 22 (May 2026): 7, <https://doi.org/10.1016/j.chbr.2026.101068>; Wong et al., "Large Scale Audiobook Synthesis," 4.
- ²² Haitao Huang et al., "A Concise Review of Long Context in Large Language Models," paper presented at the International Conference on Algorithms Software Engineering and Network Security, Nanchang, China, April 26, 2024, 563, <https://doi.org/10.1145/3677182.3677282>.
- ²³ Weicheng Zhang et al., "Audiobook Synthesis with Long-Form Neural Text-to-Speech," papers presented at the 12th ISCA Speech Synthesis Workshop, Grenoble, France, August 26, 2023, 139, <https://doi.org/10.21437/SSW.2023-22>.
- ²⁴ Wong et al., "Large Scale Audiobook Synthesis," 4.
- ²⁵ Walsh et al., "Large-Scale Automatic Audiobook Creation," 3675; Mary Burkey, "Voices in My Head," *Booklist*, June 1, 2015, 133.
- ²⁶ Shannon Maughan, "AI and Audiobooks: Another Chapter," *Publishers Weekly*, February 7, 2022, 6.
- ²⁷ Microsoft, "Project Gutenberg Brings Free Audiobooks to the Global Community."
- ²⁸ Roswitha Skare, "Paratext," *Knowledge Organization* 47, no. 6 (2020): 511, <https://doi.org/10.5771/0943-7444-2020-6-511>.
- ²⁹ Simon Rowberry, *The Early Development of Project Gutenberg c. 1970–2000*, 1st ed. (Cambridge, UK: Cambridge University Press, 2023), 27, <https://doi.org/10.1017/9781108785778>.
- ³⁰ Walsh et al., "Large-Scale Automatic Audiobook Creation," 3675; Wong et al., "Large Scale Audiobook Synthesis," 1.
- ³¹ Walsh et al., "Large-Scale Automatic Audiobook Creation," 3676.
- ³² Wong et al., "Large Scale Audiobook Synthesis," 1, 2.
- ³³ Magadán-Díaz et al., "Audiobooks and Artificial Intelligence," 407, 417; Emma Roderó and Ignacio Lucas, "Synthetic versus Human Voices in Audiobooks: The Human Emotional Intimacy Effect," *New Media & Society* 25, no. 7 (2023): 1, 3–5, 13–15, <https://doi.org/10.1177/14614448211024142>.
- ³⁴ Magadán-Díaz et al., "Audiobooks and Artificial Intelligence," 419; Ahrens-Schwabe et al., "Who Cares about Artificial Intelligence?"
- ³⁵ Kingsley, "Innovating with AI," 6:12.
- ³⁶ Walsh et al., "Large-Scale Automatic Audiobook Creation," 3676.
- ³⁷ Weber, "'Reading' the Public Domain," 210.
- ³⁸ Wong et al., "Large Scale Audiobook Synthesis," 5.
- ³⁹ Kingsley, "Innovating with AI," 6:37.
- ⁴⁰ Robert A. Carter, "Why Does a Publisher Advertise?" *Publishing Research Quarterly* 14, no. 1 (1998): 54–55, <https://doi.org/10.1007/s12109-998-0006-7>.
- ⁴¹ Seungpeel Lee et al., "What Books Will Be Your Bestseller? A Machine Learning Approach with Amazon Kindle," *The Electronic Library* 39, no. 1 (2021): 137–51, <https://doi.org/10.1108/EL-08-2020-0234>.

- ⁴² Wafaa S. El-Kassas et al., "Automatic Text Summarization: A Comprehensive Survey," *Expert Systems with Applications* 165 (March 2021): 4, <https://doi.org/10.1016/j.eswa.2020.113679>; Rada Mihalcea and Hakan Ceylan, "Explorations in Automatic Book Summarization," paper presented at the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, Czech Republic, June 2007, 380.
- ⁴³ Jothan Almeida, "Prompt Engineering: A Comparative Study of Prompting Techniques in AI Language Models," paper presented at 2025 IEEE Integrated STEM Education Conference, Princeton, New Jersey, USA, March 15, 2025, 1, 2, <https://doi.org/10.1109/ISEC64801.2025.11147384>.
- ⁴⁴ John Berryman and Ziegler, "Chapter 5: Prompt Content: Few-Shot Prompting," in *Prompt Engineering for LLMs* (O'Reilly Media, Inc., 2024), https://learning.oreilly.com/library/view/prompt-engineering-for/9781098156145/ch05.html#ch05_few_shot_prompting_1728435524681340.
- ⁴⁵ Aytuğ Onan and Hesham Alhumyani, "Assessing the Impact of Prompt Strategies on Text Summarization with Large Language Models," in *Computer Applications in Industry and Engineering*, edited by Gongzhu Hu et al. (Switzerland: Springer Nature, 2025), https://doi.org/10.1007/978-3-031-76273-4_4.
- ⁴⁶ Vladimir Geroimenko, "Key Principles of Good Prompt Design," in *The Essential Guide to Prompt Engineering* (Switzerland: Springer Nature, 2025), 17–36, https://doi.org/10.1007/978-3-031-86206-9_2.
- ⁴⁷ Kingsley, "Innovating with AI," 10:56.
- ⁴⁸ Kingsley, "Innovating with AI," 9:30.
- ⁴⁹ Koraljka Golub, "Automated Subject Indexing: An Overview," *Cataloging & Classification Quarterly* 59, no. 8 (2021): 702–19, <https://doi.org/10.1080/01639374.2021.2012311>.
- ⁵⁰ Heather Moulaison-Sandy and Zach Coble, "Leveraging AI in Cataloging: What Works, and Why?," *Technical Services Quarterly* 41, no. 4 (2024): 381, <https://doi.org/10.1080/07317131.2024.2394912>; Mary Aycock, "Prompting Generative AI to Catalog: The Promise and the Reality," *College & Research Libraries News* 86, no. 10 (2025): 423–30.
- ⁵¹ Charlene Chou and Tony Chu, "An Analysis of BERT (NLP) for Assisted Subject Indexing for Project Gutenberg," *Cataloging & Classification Quarterly* 60, no. 8 (2022): 807–35, <https://doi.org/10.1080/01639374.2022.2138666>.
- ⁵² Francisco Tadeu Gonçalves De Oliveira Foz Foz et al., "Automatic Classification of Subjects and Sustainable Development Goals (SDGs) in Documents with Generative AI: An Experience from the Unicamp Library System," *Information Technology and Libraries* 45, no. 1 (2026), <https://doi.org/10.5860/ital.v45i1.17510>.