

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28

UNITED STATES DISTRICT COURT
NORTHERN DISTRICT OF CALIFORNIA
SAN FRANCISCO DIVISION

ANDREA BARTZ, ANDREA BARTZ, INC.,
CHARLES GRAEBER, KIRK WALLACE
JOHNSON, and MJ + KJ, INC., individually and on
behalf of others similarly situated,

Plaintiffs,

v.

ANTHROPIC PBC,

Defendant.

Case No. 3:24-CV-05417-WHA

Action Filed: August 19, 2024

**DECLARATION OF MOHIT IYYER IN
SUPPORT OF DEFENDANT
ANTHROPIC PBC'S OPPOSITION TO
PLAINTIFFS' MOTION FOR CLASS
CERTIFICATION**

Judge: Honorable William H. Alsup

DECLARATION OF MOHIT IYER

I, MOHIT IYER, declare as follows:

I. QUALIFICATIONS AND ASSIGNMENT

A. Qualifications

1. My name is Mohit Iyer. I am an Associate Professor of Computer Science at the University of Maryland, College Park, and I also serve as an Adjunct Associate Professor of Computer Science at the University of Massachusetts Amherst. I received my Ph.D. in Computer Science from the University of Maryland in 2017, after which I served as an Associate Professor at the University of Massachusetts Amherst and a Young Investigator at the Allen Institute for AI (AI2). At the University of Maryland, I teach graduate-level courses in the fields of Natural Language Processing (“NLP”) and Deep Learning, while supervising Ph.D. and Master’s degree students.

2. My research is focused on NLP and machine learning, particularly improving the instruction-following abilities of large language models (“LLMs”) for long-form generation, designing robust evaluation methods for long-form and multilingual text (focusing on factuality and coherence), developing collaborative human-LLM systems for creative writing, and increasing the robustness of LLM-generated text detectors against adversarial attacks such as paraphrasing. I have extensive experience developing and evaluating LLM-based systems, including language applications such as story generation, instruction-following models, and factuality evaluation.

3. I have published over 75 papers in leading journals and at conferences such as the Association for Computational Linguistics (“ACL”), the North America chapter of the Association for Computational Linguistics (“NAACL”), the Empirical Methods in Natural Language Processing (“EMNLP”), the International Conference on Machine Learning (“ICML”), the Conference on Neural Information Processing Systems (“NeurIPS”), to name a few. As of today, per Google Scholar, I have over 28,000 citations, an h-index of 44, which means 44 of my publications have been cited over 44 times, and an i-10 index of 79, which means 79 of my publications have been cited over 10 times. My co-authored work “Deep Contextualized Word

Representations”, which introduced the foundational ELMo model, has been widely cited, including in “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding” and OpenAI’s “Improving Language Understanding by Generative Pre-Training.” My recent work includes research on long-context LLMs, watermarking of LLM outputs, factuality scoring, evaluation of hallucination in generated text, and instruction tuning for complex generation tasks. I also hold patents on dynamic memory networks and contextualized word representations.

4. My contributions to the field of machine learning have been recognized with several awards, including the National Science Foundation (“NSF”) CAREER Award (2021), the Samsung AI Researcher of the Year Award (2022), and multiple best/outstanding paper awards at top conferences such as the Association for Computing Machinery Conference on Computer and Communications Security (“ACM CCS”), the European Chapter of the Association for Computational Linguistics (“EACL”), and the NAACL.

5. I have been invited to deliver keynotes at conference workshops, including at the Knowledge Discovery and Data Mining workshop at ACM’s Special Interest Group on Knowledge Discovery and Data Mining (“SIGKDD”), ICML’s workshop on Long Context Foundation Models, and ACM’s Special Interest Group on Information Retrieval’s (“SIGIR”) workshop on Retrieval-Enhanced Machine Learning. I have also been invited to give talks at universities such as New York University (“NYU”), the Massachusetts Institute of Technology (“MIT”), and the University of Tokyo, on topics related to LLMs, retrieval-augmented generation, and long context foundation models. I also serve as a Senior Area Chair for ACL and EMNLP, and I’ve co-organized several workshops on narrative understanding, document-grounded dialogue, and conversational question answering.

6. A true and correct copy of my current curriculum vitae is attached as Exhibit A, which contains additional information about my academic background, research contributions, awards, and publication record.

B. Assignment and Material Used

7. Counsel for Anthropic has asked me to perform two sets of analyses.

8. First, I have been asked to analyze the methods for identifying books in Anthropic's datasets proposed in the Expert Declaration of Ben Y. Zhao, PhD in support of Plaintiffs' Motion for Class Certification ("Zhao Declaration" or "Zhao Decl.") (ECF No. 125-6). The datasets I have analyzed are versions of [REDACTED] [REDACTED] which I understand that Anthropic acquired from the internet (collectively, the "Internet Books Datasets"). *See* Zhao Decl. ¶ 37.

9. Second, I have been asked to perform queries of the Internet Books Datasets and Anthropic's metadata for its [REDACTED] (defined below) to identify possible books in those datasets with certain characteristics.

10. To perform my analysis, I reviewed Plaintiffs' First Amended Complaint, the Zhao Declaration and the source code for Dr. Zhao's algorithm, the datasets and metadata produced in discovery identified below, and publicly available information relating to the specific datasets at issue. A list of the materials relied upon in the preparation of my declaration is attached as Exhibit B.

11. Anthropic provided me with access to the following datasets:

- [REDACTED]¹);
- [REDACTED]) and associated metadata; and
- [REDACTED]³) and associated metadata.

12. Anthropic also provided me with the following files containing metadata associated with the following [REDACTED] datasets:

- [REDACTED]
- [REDACTED]

¹ *See* Declaration of Mycal Tucker ("Tucker Decl.") ¶ 12 n.5.

² *See* Tucker Decl. ¶ 15 n.7.

³ *See* Tucker Decl. ¶ 18 n.9.

• [REDACTED]⁴

II. SUMMARY OF OPINIONS

13. In my opinion, the methods proposed by Dr. Zhao to identify books in the Internet Books Datasets will not reliably identify books in the datasets. Dr. Zhao proposes three “alternative approaches” that he would use to identify books in the Internet Books Datasets: (i) consulting metadata, (ii) consulting [REDACTED], and (iii) a “Book Hash” method.

14. Dr. Zhao’s methodologies are not based on scientific standards that have been peer reviewed or which are widely accepted in the field of natural language processing. Instead, he has created his own methodologies for the purposes of this case.

15. Dr. Zhao did not test his methodologies at scale to determine their error rate in any of the datasets, except to try to determine if the small set of nine books written by the Plaintiffs are in the Internet Books Datasets. *See* Zhao Decl. ¶¶ 66, 68.

16. Dr. Zhao’s methodologies will result in false positives⁵ and false negatives⁶, and for each of the steps of his analysis, and as a whole, he fails to identify the rates of false positives and false negatives.

17. The Internet Books Datasets in many cases contain small fragments of books, books based on factual materials (encyclopedias, atlases, etc.), books published between 1926 and 1963, books published between 1964 and 1989, and books by authors who are likely not United States citizens.

III. ANALYSES OF BOOK IDENTIFICATION METHODS FOR THE INTERNET BOOKS DATASETS

A. Dr. Zhao’s Proposed Metadata-Based Methods of Identifying Books

18. Dr. Zhao’s first proposed approach to identify books in the Internet Books Datasets

⁴ The metadata for these [REDACTED] datasets were produced as ANT_BARTZ_000002544, ANT_BARTZ_000004894; ANT_BARTZ_000436674, and ANT_BARTZ_000436675.

⁵ A “false positive” means that Dr. Zhao’s methodology would indicate that a book is present in one of the datasets, when in fact the book *is not* present.

⁶ A “false negative” means that Dr. Zhao’s methodology would indicate that a book is not present in one of the datasets, when in fact the book *is* present.

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28

4

5
6
7
8

9

10

- 1
- 2
- 3
- 4
- 5
- 6

17
18
19
20

21
22
23
24

26
27
28

23. The Anthropic [REDACTED] Dataset comprises [REDACTED] JSONL files are files with a set of information on each line. [REDACTED]

“filename” and “text” (which would contain the text of the possible “book” itself). The following examples show Anthropic’s [REDACTED] Dataset format:

24. As can be seen, each line is composed of pairs of file names and text from possible “books”; I refer to these pairs as *records*. The “filename” field in the Anthropic [REDACTED] Dataset generally, but not always, appears to provide information about the possible title and author of the possible “book” whose text is contained in the “text” field.

25. In some instances, the Anthropic [REDACTED] Dataset file names appear to accurately and clearly describe their associated text. For example, to verify the accuracy of the first record above, I manually compared the opening paragraphs of text in the “text” field with the opening paragraphs of text in the Google Books preview of [REDACTED] and found that the texts appeared to be from the same book. However, I could not verify that the complete book is there without doing further analysis of the complete text.⁹ Further, in many instances, the file name does not identify a book, or seems to identify an incorrect book.

26. To identify instances where the Anthropic [REDACTED] Dataset file names do not accurately or clearly identify their associated text, I searched the Anthropic [REDACTED] Dataset for records with very short file names (fewer than eight characters long). It should be noted that this is just one method for identifying examples of inaccurate file names, and numerous additional

⁹ The method of comparing to Google Books also cannot be used to confirm that the entire text in the dataset is the same as the Google Books version, because Google Books contains only excerpts for many books.

approaches likely exist. Using different types of searches would likely result in finding additional incorrect file names, although I am not aware of a method that could comprehensively identify all incorrect file names.

27. I searched for short file names by removing any file extension portions of file names¹⁰, and then programmatically identifying file names shorter than 8 characters. Searching for short file names yielded 335 file names with fewer than 8 characters. If a file name is very short, it is highly unlikely to be useful to identify the book, as it would be unlikely to reflect author, title, or even ISBN information (which is typically 10 or 13 numerical digits) or ASIN information (Amazon Standard Identification Numbers are 10 character alphanumeric identifiers).

28. I reviewed samples of the identified results and confirmed that this method identified numerous examples of records whose file names do not accurately or clearly identify their associated texts. The following tables list ten of these example records, with file names compared to what I believe may be the correct titles and authors associated with the content of their text records. To identify the possibly correct titles and authors of the books in the text fields, I manually examined the text content and compared it with Google Books previews as before. The file names in Anthropic's Dataset do not appear to correspond to correct or complete information about book title or author, at least in some cases.

Short File Names

		Inferred correct title and author
1		<i>Mathematics, Substance and Surmise: Views on the Meaning and Ontology of Mathematics</i> by Ernest Davis and Philip J. Davis
2		<i>1861 : The Civil War Awakening</i> , by Adam Goodheart
3		<i>Kali Linux – Assuring Security by Penetration Testing</i> , by Lee Allen, Tedi Heriyanto, and Shakeel Ali

¹⁰ Many end with “.epub.txt” because the texts come from epub (eBook) files that have been converted to text files.

4	[REDACTED]	<i>Agile Data Science: Building Data Analytics Applications with Hadoop</i> by Russel Journey
5	[REDACTED]	<i>Kierkegaard's World (How to Believe)</i> by Clare Carlisle
6	[REDACTED]	<i>What?: Are These the 20 Most Important Questions in Human History--Or is This a Game of 20 Questions?</i> by Mark Kurlansky
7	[REDACTED]	<i>Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More</i> by Matthew Russell
8	[REDACTED]	<i>Shopify Application Development: Build highly effective Shopify apps using the powerful Ruby on Rails framework</i> by Michael Larkin
9	[REDACTED]	<i>High Performance Drupal: Fast and Scalable Designs</i> by Jeff Sheltren, Narayan Newton, Nathaniel Catchpole
10	[REDACTED]	<i>Curious: The Desire to Know and Why Your Future Depends On It</i> by Ian Leslie

b. [REDACTED]

29. Dr. Zhao also references a “[REDACTED]” as a potential source of metadata to identify the books in Anthropic’s [REDACTED] Dataset. Zhao Decl. ¶ 50. But [REDACTED] is not a reliable way to determine what is in Anthropic’s version of [REDACTED] because: (i) [REDACTED] admits the [REDACTED] is unreliable, (ii) [REDACTED] did not disclose the methodology it used to create the [REDACTED] with sufficient detail to allow validation of its accuracy, and (iii) the version of [REDACTED] is likely not the same as the Anthropic [REDACTED] Dataset.

30. First, [REDACTED] admits the [REDACTED] is unreliable. [REDACTED]

1 [REDACTED]

2 [REDACTED]

3 [REDACTED]

4 [REDACTED]

5 [REDACTED]

6 [REDACTED]

7 31. [REDACTED] also admits that he was unable to find any identifying information for

8 thousands of books in [REDACTED]. When he first disclosed th [REDACTED], he admitted that out of

9 [REDACTED] in the dataset, “[REDACTED]

10 [REDACTED]”¹² In a [REDACTED]

11 [REDACTED]

12 [REDACTED]¹³

13 32. Second, [REDACTED] does not disclose its methodology with sufficient detail to

14 validate its accuracy. [REDACTED] describes the method he used to create [REDACTED] at a very high level

15 as follows:¹⁴

- 16 • [REDACTED]
- 17 [REDACTED]
- 18 • [REDACTED]
- 19 [REDACTED]
- 20 • [REDACTED]
- 21 [REDACTED]

22

23 [REDACTED]

24 [REDACTED]

25 [REDACTED]¹²

26 [REDACTED]

27 [REDACTED]

28 [REDACTED]¹³

[REDACTED]¹⁴

1 • [REDACTED]
2 [REDACTED]
3 • [REDACTED]
4 [REDACTED]
5 [REDACTED]
6 33. Third, even if one assumes that [REDACTED] used scientifically valid techniques and/or
7 that the error rate resulting from his methodology was acceptable, [REDACTED]
8 would still not be a reliable way to determine whether a book is present in the Anthropic
9 Dataset because Anthropic likely possesses a different version of [REDACTED] than the one [REDACTED]
10 [REDACTED] has. I understand that Anthropic acquired its [REDACTED] Dataset around February 2021,
11 whereas [REDACTED] states that he acquired his version of [REDACTED] in the summer of 2023. Tucker
12 Decl. ¶ 9. Dr. Zhao has not disclosed any analysis to demonstrate that these versions—
13 downloaded years apart from each other—contain the same data. In fact, according to Dr. Zhao,
14 Anthropic's [REDACTED] Datasets comprises [REDACTED] records, whereas [REDACTED]
15 [REDACTED] records.¹⁵

16 34. Furthermore, Anthropic's [REDACTED] Dataset could be different from [REDACTED] due
17 to extraction errors that occurred in connection with Anthropic's download, or the creators'
18 compilation of the [REDACTED] Dataset. However, there is no definitive way to know whether these
19 extraction errors are present in the original version of [REDACTED] or in [REDACTED]
20 [REDACTED] version without actually comparing those datasets.

21 35. I examined data in the Anthropic [REDACTED] Dataset to identify text extraction errors,
22 which are errors resulting from failed or incomplete extractions of eBook text. These errors result
23 in missing content or truncated texts in the "text" fields of Anthropic's [REDACTED] Dataset records. I
24 identified two types of extraction errors: full and partial extraction error files. Full extraction error
25 files are completely empty, containing no content at all, while partial extraction error files include
26 minimal content but are effectively empty (e.g., containing only a Table of Contents). To identify
27

28 ¹⁵ [REDACTED].

potential candidates in these categories, I developed a script that applies a predefined set of conditions to flag relevant files.

36. The script uses the following criteria to flag extraction error files:

- The file is a full extraction error file if, and only if, the byte size of the file is 0 bytes.
- The file is a partial extraction error file if:
 - The byte size of the file is less than 10,000 bytes, and
 - The average line length is 20 characters or fewer.

37. Using these criteria, I obtained the following results for the Anthropic Dataset:

- Full extraction error files: 1,284
- Partial extraction error files: 456

38. A manual review of identified full extraction error files confirmed that the identified Anthropic Dataset entries are empty, confirming that no text was extracted from these books.

39. A manual review of the flagged partial extraction error files was conducted to identify any false positives, meaning short files that did not actually reflect extraction errors. Out of 456 partial extraction error files, only 14 were determined to be false positives. These appeared to be primarily poetry files and short children's stories, which naturally contain shorter and fewer lines, making them match the defined criteria despite being valid files.

40. The criteria for identifying partial extraction error files could be broadened (e.g., by increasing the maximum file size or line length) to capture more true positives. However, this approach would likely also increase false positives that would need to be filtered with manual review. As a result, while 442 partial extraction error files have been confirmed in Anthropic's Dataset (out of 1,726 total extraction error files), expanding the criteria would almost certainly reveal additional cases.

41. In sum, to the extent Dr. Zhao suggests using [REDACTED] to identify books in Anthropic's Dataset, that would not be a scientifically valid approach to

identify books in the Anthropic [REDACTED] Dataset.

2. [REDACTED] - Third Party Metadata Approach

42. Dr. Zhao states that the Anthropic [REDACTED] Datasets (the other two Internet Books Datasets) “*often* have metadata associated with the underlying dataset that records *certain* information regarding the books present.” Zhao Decl. ¶ 50 (emphasis added). However, Dr. Zhao does not state how “often” this occurs, what the “certain” available metadata is, or make any attempt to prove that this metadata is complete or reliable. Moreover, in my opinion, the metadata in these sources is far from complete or reliable.

43. As with Anthropic’s [REDACTED] Dataset, Dr. Zhao proposes two metadata-based approaches to identifying books in the Anthropic [REDACTED] Dataset. First, he proposes relying on the metadata that was compiled by the third parties that created [REDACTED] Zhao Decl. ¶ 50. Second, he proposes relying on a [REDACTED]. For the Anthropic [REDACTED] Dataset, Dr. Zhao only proposes relying on the metadata that was compiled by the third parties that created [REDACTED]. *Id.*

a. The Anthropic [REDACTED] Metadata

44. According to public reporting, [REDACTED]

¹⁶ Carolyn Caffrey Gardner & Gabriel J. Gardner, *Fast and Furious (at Publishers): The Motivations behind Crowdsourced Research Sharing*, 78 College & Research Libraries 131, 132 (2017), available at https://www.researchgate.net/publication/317117356_Fast_and_Furious_at_Publishers_The_Motivations_behind_Crowdsourced_Research_Sharing.

¹⁷ *Id.*

¹⁸ [REDACTED]

¹⁹ [REDACTED]

45. The Anthropic [REDACTED] Dataset is separated into [REDACTED] portions, and is composed of records in a variety of different file formats, including “epub”, “pdf”, and “cbz.” I counted [REDACTED] in the [REDACTED] portion and [REDACTED] in the [REDACTED] portion, each a single file.

46. The Anthropic [REDACTED] Dataset is also composed of records in a variety of different file formats, including “txt”, “pdf”, and “epub.” I counted [REDACTED] records, each a single file.

47. Unlike Anthropic’s [REDACTED] Dataset, there is some structured metadata associated with Anthropic’s [REDACTED] Datasets. However, the metadata in these datasets has been crowd-sourced over years²⁰ and is extremely prone to errors. Because of the interrelated origins of [REDACTED] and [REDACTED], they share common metadata failings.

48. A comment on a YCombinator.com forum describes the metadata creation process for [REDACTED] as follows:

[REDACTED]

49. A forum on [REDACTED] allows reporting of [REDACTED] URLs that point to a book that does not match the corresponding metadata.²² There are [REDACTED] While the thread indicates some of the reported errors were fixed, the high rate at which new posts have been made over the years indicates that the book/metadata mismatch is a widespread problem with [REDACTED].

50. I searched for instances where Anthropic’s [REDACTED] metadata records do not

²⁰ [REDACTED]

²¹ [REDACTED]

²² [REDACTED]

accurately describe their associated possible “book” files by reviewing posts in the [REDACTED] for instances where users described errors in other [REDACTED] metadata.

51. After identifying posts listing known (or at least suspected) metadata errors, I identified the noted records in the Anthropic [REDACTED] Dataset by locating the Anthropic [REDACTED] metadata record associated with the md5 hash²³ listed in the post. I repeated for the latest ten forum posts identifying errors in [REDACTED] records added to [REDACTED] before 2022.²⁴ I then compared the possible “book” file associated with the metadata record with the metadata and found that in all ten cases the metadata attributes described a completely different book from the possible “book” in the metadata record’s associated raw book file. For example, the metadata record for the [REDACTED] record for [REDACTED] is associated with a “book” file containing the text of [REDACTED]

²⁵

52. The following table lists those ten instances:

	[REDACTED]			[REDACTED]
	Title	Author	Publication Year	
1	<i>Clueless in the Kitchen: A Cookbook for Teens</i>	Evelyn Raab, George A. Walker	1998	[REDACTED]
2	<i>Set Theory</i>	Jill B. Becker, Karen J. Berkley, Nori Geary, Elizabeth	2003	[REDACTED]

²³ An **md5 hash** is an alphanumeric identifier used in the [REDACTED] database to uniquely identify books. It is generated by passing the raw binary contents of a file through MD5, a cryptographic hash function.

²⁴ I limited my search to books added before 2022 to increase the likelihood that these books would be present in Anthropic’s [REDACTED] dataset.

²⁵ This was confirmed by examining the cover of the file, which listed this title and authors. I further compared the opening paragraph of the book file with the Google Books preview for [REDACTED] and verified that they matched.



		Hampson, James P. Herman, Elizabeth Young		
3	<i>Philosophy Guidebook to Kant and the Critique of Pure Reason</i>	V. Spike Peterson	1998	
4	<i>Handbook of Reagents for Organic Synthesis</i>	Philip L. Fuchs	2013	
5	<i>Python and Tkinter Programming</i>	John E. Grayson	2000	
6	<i>Compact cabins : simple living in 1,000 square feet or less</i>	Gerald Rowan	2009	
7	<i>The Lost Science of Money: The Mythology of Money, The Story of Power</i>	Stephen A. Zarlenga	2002	
8	<i>The Tibetan Book of Living and Dying</i>	Sogyal Rinpoche, Patrick D. Gaffney, Andrew Harvey	1994	
9	<i>1.e4 vs The French, Caro-Kann & Philidor</i>	Parimarjan Negi	2014	

10	<i>Indicators of Equal Employment Opportunity: Status and Trends</i>	Rosa Luxemburg	2003	
----	--	----------------	------	--

53. I also identified errors in the [REDACTED] metadata by comparing [REDACTED] metadata records with their associated raw possible book files. Many of the [REDACTED] raw book files referenced in its metadata are not present in the actual dataset. For example, the first book in the [REDACTED] table above, *Clueless in the Kitchen: A Cookbook for Teens*, by Evelyn Raab and George A. Walker, has an entry in the [REDACTED] metadata, and its metadata [REDACTED] [REDACTED], the indicated raw book file does not exist in the [REDACTED] dataset, representing a significant discrepancy between the metadata record and its raw book content. In total, while only [REDACTED] raw book files exist in [REDACTED] there are at least [REDACTED] entries²⁶ in the [REDACTED] metadata, and thus less than 30% of [REDACTED] metadata entries are associated with a book file. The following table lists results from searching for the books in the [REDACTED] metadata errors table above in [REDACTED]. While all of the books (as identified by their unique hash identifier) were present in the [REDACTED] metadata, none of their associated book files were present in the dataset.

					Book in Dataset
	Title	Author	Publication Year		
1	<i>Clueless in the Kitchen: A Cookbook for Teens</i>	Evelyn Raab, George A. Walker	1998		Not Present

²⁶ There are a total of [REDACTED] rows in the [REDACTED] metadata file, but many of the rows are not valid metadata entries (e.g., additional columns, incorrectly formatted rows). Using pandas, the Python library, to load in the metadata file and ignore bad lines results in [REDACTED] entries.

2	<i>Set Theory</i>	Jill B. Becker, Karen J. Berkley, Nori Geary, Elizabeth Hampson, James P. Herman, Elizabeth Young	2003		Not Present
3	<i>Philosophy Guidebook to Kant and the Critique of Pure Reason</i>	V. Spike Peterson	1998		Not Present
4	<i>Handbook of Reagents for Organic Synthesis</i>	Philip L. Fuchs	2013		Not Present
5	<i>Python and Tkinter Programming</i>	John E. Grayson	2000		Not Present
6	<i>Compact cabins : simple living in 1,000 square feet or less</i>	Gerald Rowan	2009		Not Present
7	<i>The Lost Science of Money: The Mythology of Money, The Story of Power</i>	Stephen A. Zarlenga	2002		Not Present
8	<i>The Tibetan Book of Living and Dying</i>	Sogyal Rinpoche, Patrick D. Gaffney, Andrew Harvey	1994		Not Present
9	<i>1.e4 vs The French, Caro- Kann & Philidor</i>	Parimarjan Negi	2014		Not Present
10	<i>Indicators of Equal Employment Opportunity:</i>	Rosa Luxemburg	2003		Not Present

1					
2	Status and Trends				

b.

54. Similar to what it did for

which Dr. Zhao states he may rely upon to identify books in Anthropic's version of

55. This suffers from the same types of inadequacies as

56. First, gain acknowledges that the is error-prone and prominently disclaims correctness:

57. Second, does not disclose any information about the methodology it used to create this so it is not possible to scientifically validate its accuracy.

58. Third, is also almost certainly different from the Anthropic Dataset. I understand that Anthropic obtained its Dataset around June 2021. However, states that its version of the dataset is a . This is significant because the content of is a crowd-sourced dataset that is constantly changing, such that looking at the content of one version created this year is an entirely unreliable way to determine the contents of a version obtained four years ago.

59. Similar to the Anthropic Dataset, there is also evidence of errors in the extraction, uploading, and/or downloading process for the Anthropic Datasets. I examined raw book files in the and datasets to identify instances where the raw book file is fragmented or malformed. To search for these instances, I identified the smallest EPUB (eBook) and PDF files in the Dataset and the smallest overall files in the

27

Dataset.

60. I manually reviewed the five smallest EPUB and PDF files in the Dataset and five smallest overall files in the dataset. All files displayed significant errors: the PDF files were empty, the EPUB files could not be loaded or viewed as eBooks, and the smallest files contained malformed error messages. These errors likely occurred while the or datasets were uploaded, downloaded, or extracted.

61. Indeed, as noted with respect to any attempt to compare its database with a version

Therefore, there is also every reason to believe that includes numerous books that are not present in

B. Dr. Zhao's "Book Hash" Approach

62. As an alternative to his proposal to use third party metadata or to identify possible books in the Internet Books Datasets, Dr. Zhao proposes a Book Hash approach that relies on "one-way hashing algorithms," or "hash codes." Zhao Decl. ¶ 54. Hash codes are commonly used for matching texts that are exactly the same, but are imperfect for fuzzy matching, particularly in the way that Dr. Zhao applied his.

63. It is not possible to use Dr. Zhao's hash method to comprehensively identify all of these millions of books because the hash method presupposes the existence of a "source of truth," an authoritative digitized version of each book that might be in a given dataset. But Dr. Zhao does not specify where such "source of truth" copies would be obtained. Separately, Dr. Zhao's methodology will be unreliable for several reasons, meaning it will undeniably result in false positives and false negatives. He does not attempt to identify the rates of false positives and false negatives associated with his method.

²⁸ *Id.*

64. ***Dr. Zhao Fails to Identify a Source of Truth Dataset.*** First, the hash method is not feasible to implement in reality because it requires a “source of truth” and Dr. Zhao fails to identify how the parties would obtain one. Put simply, hashing is a *matching* technology. Dr. Zhao’s proposed method requires comparing every sentence with more than ten words from an authoritative digitized copy of each book against Anthropic’s datasets. Zhao Decl. ¶¶ 60, 63. If there is a sufficiently high degree of match, Dr. Zhao would declare it a positive identification. *Id.* ¶ 64.

65. To use this method, one must have a digitized copy of every book that *may* be in Anthropic’s datasets. For example, even if Dr. Zhao had digitized copies of all 200,000 books in the HarperCollins catalogue, he could not use the hash method to rule any book in the Penguin Random House catalogue in or out of a dataset. In short, using Dr. Zhao’s hash method to determine the contents of a dataset would require obtaining a digitized copy of each of the many millions of books that may or may not be in an Anthropic dataset.

66. But Dr. Zhao provides no indication of where these authoritative copies will come from, creating a fundamental, practical flaw in his proposal. Because Dr. Zhao has no plan to obtain all such authoritative copies of books, his hash method cannot be used to ascertain the contents of a dataset.

67. ***Dr. Zhao’s Book Hash Method Is Unreliable and He Does Not Address the Numerous False Positives and False Negatives That It Would Produce.*** Dr. Zhao’s hash method requires the determination of several thresholds, and setting those thresholds has consequences on the results: setting too strict a threshold means missing valid matches (false negatives) and setting too permissive a threshold means finding invalid matches (false positives). Finding the optimal setting often requires tuning and validation, and in any event requires making calibration choices. As an initial matter, Dr. Zhao sets several important thresholds for the hash comparison based on his subjective judgment, without explaining his reasoning. These choices include setting the minimum sentence length to be analyzed at 10 words or more and [REDACTED] [REDACTED] Moving the threshold for when to ignore a sentence up or down will change whether a particular snippet of a book is considered a match. There are many

1 sentences of more than ten words that appear in many books. For example, this quote from
2 Ecclesiastes appears littered through literature and contains more than ten words: “To every thing
3 there is a season, and a time to every purpose under heaven.” A change in this particular parameter
4 will result in large differences in what is considered or not considered a match.

5 68.

6
7 29

8
9
10 30

11 This

12 arbitrary change is a further indicator of the unreliable and untested nature of Dr. Zhao’s hash
13 method.

14 69.

15
16 Since many potential

17 “book” records will include boilerplate text and identifying information from the publisher at the
18 beginning, this is particularly important, and

19
20
21 Because of this, it is likely that there will be many false positive matches. The hash
22 method does not require that sentences be in the same order or even in the same location in the
23 two books under consideration, greatly increasing the likelihood of a false positive.

24 70. Dr. Zhao also proposes a validation method in which the books are divided into
25 equal subsets and the subsets are then analyzed to determine whether the matching rate for each

26
27 ²⁹ Dr. Zhao’s source code was produced as BARTZ_ZHAO_CERT_000000506 and
BARTZ_ZHAO_CERT_000000507.

28 ³⁰ Dr. Zhao’s later-produced source code was produced as BARTZ_ZHAO_CERT_000000509.

subset is “approximately” equal to the matching rate for the whole book in order to determine whether a full copy of the book is present. Zhao Decl. ¶ 64. [REDACTED] Moreover, Dr. Zhao does not define what it means for rates to “approximately” match. Making the tolerance stricter (such that the rate from each subset must be very close to the overall rate) would reduce incorrect matches, but would also possibly eliminate some correct matches. On the other hand, lowering the tolerance would make incorrect matches more likely, but would reduce the risk of missing correct matches.

71. Furthermore, this algorithm is fundamentally flawed. For example, Dr. Zhao’s method could suggest dividing a book into two equal parts—i.e., the subsets. Assume that each half has 10 sentences. Then assume two sentences in each half are a match for the copy provided. In that case, the matching rate for the subsets (2 out of 10 for each) would be the same as the matching rate for the whole (4 out of 20), and Dr. Zhao would consider this evidence of the entire book being present in the copy, not a mere quotation. But it is clear that quoting two sentences does not imply that “the book—rather than a summary or lengthy quotation—is present within the data.”

72. Dr. Zhao’s method is also very sensitive to formatting differences. For example, if one copy of a book has line breaks mid-word in many sentences but another copy does not have these word breaks, the hashing algorithm will deem them to be different, even if a human would understand them to be duplicates. For example, suppose the source-of-truth book is a 2020 edition of a certain book and a 2025 edition of the same book exists in one of Anthropic’s datasets. Because the 2020 edition of the book in question may not have the same formatting as the 2025 edition, and because Dr. Zhao’s hash method is so sensitive to formatting differences (as described below), Dr. Zhao’s method is likely to overlook this match, based on formatting differences.

73. For example, Dr. Zhao’s method handles newlines and tabs in a way that is likely to lead to errors. [REDACTED]

74. Dr. Zhao testified at his deposition that [REDACTED]

[REDACTED] Declaration of Douglas A. Winthrop in Support of Defendant Anthropic PBC's Opposition to Plaintiffs' Motion for Class Certification, Ex. 40, Deposition of Ben Y. Zhao ("Zhao Dep.") at 79:23-80:3. [REDACTED]

The table below shows examples of this behavior, using the sentence "She wandered through the quiet forest, listening to birdsong and savoring the crisp morning air." To generate the hashes in the table below, I used [REDACTED], without making any modifications. As shown in the table, [REDACTED]

Sentence	Changes	Hash
She wandered through the quiet forest, listening to birdsong and savoring the crisp morning air.	Base sentence	850174d76e82d8b90cafd4e4fe6a7db007d73b3ee00f32e041391dd645d9003b
She wandered through the quiet forest, listening to birdsong and savoring the crisp morning air.	Multiple spaces between "wandered" and "through", and "forest," and "listening"	1dba269db1424a6afb64979a7d9ff0c563e8cb7afa41fe5c9b0e9792c9eac62b
She wandered through the quiet forest, listening to birdsong and savoring the crisp morning air.	Newline characters between "wandered" and "through", and "forest," and "listening"	1dba269db1424a6afb64979a7d9ff0c563e8cb7afa41fe5c9b0e9792c9eac62b

75. Other areas where similar problems are likely to occur include: errors in line and paragraph spacing, chapter titles and headers mixed in with body text, how to handle conversion of special characters (which are particularly common in academic, scientific, mathematics, and technical books), and footnotes appearing out of place.

76. These formatting differences would be amplified to the extent that different conversion technologies were used to create the source of truth and the Internet Books Datasets, which is inevitable. To perform the hash algorithm, Dr. Zhao must first create a .txt file to use as a source of truth to compare against Anthropic's datasets. He does not say if his intent is to start with physical books or ebooks. For physical books, Dr. Zhao would have to scan the book using optical character recognition ("OCR") software, and then re-format it into a .txt file. For ebooks (books in digital format), there are many different possible file formats, such as .epub. These files would also need to be converted to .txt files as part of the hash methodology, which could require using different conversion technologies. The Internet Books Datasets themselves were created by third parties, also presumably using numerous different methods. The heterogenous ways in which the Internet Books datasets were created and/or converted to .txt files are likely to create significant problems for Dr. Zhao's hash algorithm.

77. The last step of Dr. Zhao's hash method is "a final robustness check may be performed which simply involves viewing a portion of Anthropic's data identified alongside the book in question to verify the presence of the book in Anthropic's data." Zhao Decl. ¶ 65. Put more simply, the last step of Dr. Zhao's hash method is for a human being to view the two files under consideration side-by-side to make a subjective determination if they reflect the same book.

78. According to Dr. Zhao, the total number of candidate files that needs to be compared may be as many as [REDACTED] possible records. Zhao Decl. ¶ 20 [REDACTED], ¶ 24 [REDACTED], ¶ 33 [REDACTED]. Dr. Zhao's robustness step is not practical:

- He does not say what would qualify as a sufficiently robust human review;
- Human review will be inherently subjective and error-prone; and

- Human review will be extraordinarily time consuming.

79. Normally, when an individual proposes a new text matching algorithm, it would be industry standard to test and validate that algorithm works before deploying it.

Zhao Dep. at 84:18-85:15. Simply spot checking whether a method works for an unspecified number of examples is not an adequate or reliable way to validate an algorithm.

IV. THE CONTENTS OF ANTHROPIC'S BOOKS DATASETS

80. I performed analyses to identify records with the following characteristics in the Internet Books and Anthropic Datasets: (1) fragments of books; (2) factual books; (3) books published between 1926-1963; (4) books published between 1964-1989; and (5) books by authors who are likely not United States citizens.

81. As discussed in detail above, there is inconsistent availability and reliability of metadata across the three Internet Books Datasets. Because of this, I used different methods to identify and confirm the presence of particular books in each dataset.

82. To search for possible evidence of books in the Anthropic Anthropic Dataset, I searched for file names that contained an apparent title of the book. I then manually compared the opening paragraphs of text in the "text" field with the opening paragraphs of text in the Google Books preview of the book indicated by the file name to confirm that the text appeared consistent with the file name.³¹

83. To search for possible evidence of books in the Anthropic Anthropic Dataset, I consulted the table of metadata associated with Anthropic's Anthropic Dataset. This table included a column that appeared to contain the titles of books. I first searched for the book's title in this column. Each row of the Anthropic Anthropic metadata table also contained a file location for the possible "book" record associated with the row. If at least one row's title field contained the title

³¹ As noted above, the method of comparing to Google Books cannot be used to confirm that the entire text in the dataset is the same as the Google Books version, because Google Books contains only excerpts for many books.

[REDACTED]

1 of the book, I examined the possible “book” record for that row and, similar to the procedure
2 described above for Anthropic’s [REDACTED] Dataset, manually compared the text of the book to the
3 Google Books preview to determine if it appeared consistent with the metadata.

4 84. To search for possible evidence of books in the Anthropic [REDACTED] Dataset, I
5 consulted the table of metadata associated with the [REDACTED] Dataset. This table included a column
6 that appeared to contain the titles of books, and a second column that appeared to contain names
7 of authors. I first searched for the book’s title and author in these columns. Each row of the
8 [REDACTED] metadata table also contained a file location for the possible “book” file associated with
9 the row. If at least one row’s title and author fields contained the title and author of the book, I
10 examined the possible “book” record for that row³² and, similar to the procedures described above
11 for [REDACTED] and [REDACTED], manually compared the text to the Google Books preview to determine if
12 it appeared consistent with the metadata.

13 85. To search for possible evidence of books in the [REDACTED] Dataset, I searched
14 the [REDACTED] Dataset metadata for records whose titles matched the book’s title. I then
15 reviewed any such metadata records to verify that the record’s ISBN matched the ISBN for at least
16 one version of the book, and that the record’s listed author was the author of the book.

17 **A. The Internet Books Datasets contain many records that do not appear to be**
18 **books or appear to be small fragments of books.**

19 86. I searched for book files in the Internet Books Datasets that either do not contain
20 books or contain only small fragments of books.

21 87. To search for book files in [REDACTED] that do not contain books or contain only small
22 fragments of books, I identified all files in [REDACTED] with less than 10,000 characters. For reference,
23 there are about 5 characters in the average English word and about 2,000 characters per page.³³ I

24
25 ³² In all cases, I used the [REDACTED] (ANT_BARTZ_000436666, as did Zhao (*see*
Zhao Decl. ¶33)).

26 ³³ See Vladimir V. Bochkarev et al., *The Average Word Length Dynamics as an Indicator of*
27 *Cultural Changes in Society*, 12 Soc. Evolution & Hist. 153, 155 (2015), available at
28 https://www.sociostudies.org/journal/files/seh/2015_2/153-175.pdf; Don Leeper, *How many pages*
will your book have? A castoff can tell you, Bookmobile (July 13, 2015), available at
<https://www.bookmobile.com/print-positive/how-many-pages-will-your-book-have/>.

counted the number of characters in the text field for each row of the [REDACTED] dataset, then sorted all counts according to the number of characters. The below table shows the number of records in [REDACTED] containing fewer than 10, 100, 200, 500, 1,000, 2,000, 5,000, and 10,000 characters.

	10 char.	100 char.	200 char.	500 char.	1000 char.	2000 char.	5000 char.	10,000 char.
[REDACTED]	[REDACTED]	[REDACTED]	[REDACTED]	[REDACTED]	[REDACTED]	[REDACTED]	[REDACTED]	[REDACTED]

88. I reviewed a sample of these results and found that many of them included entirely empty records or clear instances where the record contained only a small fragment of a possible book. For example, many records contained only the table of contents of a book. As book texts get longer, they are more likely to contain entire books, but my review indicated that files with up to 10,000 characters often still contain only a fragment of the book.

89. To search for book files in [REDACTED] that do not contain books or contain only small fragments of books, I identified raw book files in [REDACTED] that are small in file size. I first measured the file size of each raw book file, then grouped the book files by extension (so all PDF files were grouped together, for example), and then examined the five smallest PDF and five smallest EPUB files in the [REDACTED] dataset. I found that all ten of these files did not contain a book, or even a fragment of a book. The following table lists the files examined, their listed contents in the [REDACTED] metadata, and their actual contents.

	[REDACTED]	[REDACTED] Metadata Entry		Actual Content
		Title	Author	
1	[REDACTED]	[REDACTED]	[REDACTED]	Blank PDF File
2	[REDACTED]	[REDACTED]	[REDACTED]	Blank PDF File
3	[REDACTED]	[REDACTED]	[REDACTED]	Blank PDF File
4	[REDACTED]	[REDACTED]	[REDACTED]	Blank PDF File

5				Blank PDF File
6				1-line text file
7				3-line text file
8				Malformatted Epub File
9				Malformatted Epub File
10				Malformatted Epub File

90. To search for book files in [REDACTED] that do not contain books or contain only small fragments of books, I followed a similar process. Again, I first measured the file size of each raw book file, then examined the five smallest raw book files³⁴ that contained at least one byte in the [REDACTED] dataset and included author and title fields in the metadata. I found that all five of these files did not contain a book, or even a fragment of a book. The following table lists the files examined, their listed contents in the [REDACTED] metadata, and their actual contents.

		Metadata Entry		Actual Content
		Title	Author	
1				1-line text file
2				2-line text file
3				1-line text file

³⁴ [REDACTED] files are not identified by file extension, so I did not group files by extension and instead examined the 5 absolute smallest files across the entire [REDACTED] dataset.

4				2-line text file
5				2-line text file

B. The datasets appear to have many books composed of factual works, such as textbooks.

91. The Internet Books Datasets and [REDACTED] Datasets include many records that appear to contain text from factual works, such as textbooks.

92. To identify examples of factual works in each of the datasets, I queried each dataset to find keyword matches. I searched for the word “encyclopedia,” “atlas,” or “textbook” in the first 2,000 words of each document for the Anthropic [REDACTED] Dataset and in Anthropic’s metadata for the [REDACTED] Dataset and [REDACTED] Dataset and the [REDACTED] Dataset, and saved keyword matches as candidate factual works. Finally, I manually inspected these candidates to identify true positives. The following paragraphs provide additional details and list the results of these searches.

	Author	Title	Factual Work Keyword	[REDACTED]	[REDACTED]
1	W. Ellenberger and Francis A. Davis	An Atlas of Animal Anatomy for Artists	Atlas	[REDACTED]	[REDACTED]
2	Dan M. Fliss and Ziv Gil	Atlas of Surgical Approaches to Paranasal Sinuses and the Skull Base	Atlas	[REDACTED]	[REDACTED]
3	Thomas Suárez	Early Mapping of the Pacific	Atlas	[REDACTED]	[REDACTED]
4	Nick Kanas	Start Maps: History, Artistry,	Atlas	[REDACTED]	[REDACTED]

		and Cartography			
5	John Nolte	The Human Brain: An Introduction to Its Functional Anatomy	Atlas		
6	Vincent Terrace	Encyclopedia of Television Shows	Encyclopedia		
7	Roland Greene, Stephen Cushman, Clare Cavanagh, Jahan Ramazani, and Paul Rouzer	The Princeton Encyclopedia of Poetry and Poetic	Encyclopedia		
8	Denise Cush, Catherine Robinson, and Michael York	Encyclopedia of Hinduism	Encyclopedia		
9	Everett Ferguson	Encyclopedia of Early Christianity	Encyclopedia		
10	Charles Russell Coulter and Patricia Turner	Encyclopedia of Ancient Deities	Encyclopedia		
11	Insun Lee	Elementary Korean Workbook: A Complete	Textbook		

		Language Activity Book for Beginners			
12	Scott R. Steele, Tracy L. Hull, Neil Hyman, Justin A. Maykel, Thomas E. Read, and Charles B. Whitlow	The ASCRS Textbook of Colon and Rectal Surgery	Textbook		
13	Walter Dittrich	Classical and Quantum Dynamics -- From Classical Paths to Path Integrals	Textbook		
14	Vladimir A. Zorich	Mathematical Analysis I	Textbook		
15	Nancy deClaiss-Walford	Biblical Hebrew: An Introductory Textbook	Textbook		

	Author	Title	Factual Work Keyword		
1	V. Suzanne Klimberg	Anatomy: A Photographic Atlas	Atlas		
2	Johannes W. Rohen Chihiro Yokochi, Elke Lutjen-Drecoll	Atlas of the Histology of Brain Tumors	Atlas		
3	K.J. Zulch	Atlas of Pediatric	Atlas		

		Cutaneous Biodiversity: Comparative Dermatologic Atlas of Pediatric Skin of All Colors			
4	Nanette B. Silverberg	Bali Street Atlas Third Edition: Bali's Most Up-To-Date Street Atlas	Atlas		
5	V. Suzanne Klimberg, Courtney Townsend	Atlas of Breast Surgical Techniques	Atlas		
6	Frank K. Flinn	Encyclopedia of Catholicism	Encyclopedia		
7	Gale Group, Rosalyn Carson-Dewitt	Encyclopedia of Drugs, Alcohol, and Addictive Behavior	Encyclopedia		
8	Robert U Ayres, Robert Costanza, Jose Goldemberg, et al.	Encyclopedia of Energy Volume I	Encyclopedia		
9	Richard C. Selley, Robin Cocks, Ian Plimer	Encyclopedia of Geology Volume 1	Encyclopedia		
10	Alan V. Murray	The Crusades: An Encyclopedia	Encyclopedia		
11	James L. Levenson	The American Psychiatric Publishing Textbook of Psychosomatic Medicine	Textbook		

12	John Block, John M. Beale	Wilson & Gisvold's Textbook of Organic Medicinal and Pharmaceutical Chemistry	Textbook		
13	Tadataka Yamada, David H. Alpers, Anthony N. Kalloo, Neil Kaplowitz, Chung Owyang, Don W. Powell	Textbook of Gastroenterology, 2 Volume Set, 5th Edition	Textbook		
14	Myron F. Weiner, Anne M. Lipton	American Psychiatric Publishing Textbook of Alzheimer's Disease and Other Dementias: The APP Textbook of Geriatric Psychiatry Diagnostic Issues in Dementia	Textbook		
15	Howard M. Fillit MD, Kenneth Rockwood, Kenneth Woodhouse	Brocklehurst's Textbook of Geriatric Medicine and Gerontology, Seventh Edition: Expert Consult	Textbook		

	Author	Title	Factual Work Keyword	
1	Joshua Foer	Atlas Obscura	Atlas	
2	Lofi J.,	Seismic Atlas of the	Atlas	

	Deverchere J., Gaullier V., et al.	Messinian Salinity Crisis markers in the Mediterranean and Black Seas		
3	Nathan M. Slovis DVM Dipl ACVIM	Atlas of Equine Endoscopy	Atlas	
4	Jesse Reiser	Atlas of Novel Tectonics	Atlas	
5	Siba P. Dubey, Charles P. Molumi	Color Atlas of Head and Neck Surgery: A Step-By-Step Guide	Atlas	
6	Dorling Kindersley	Space: a Visual Encyclopedia	Encyclopedia	
7	Theresa Cheung	The Element Encyclopedia of 20,000 Dreams	Encyclopedia	
8	Martha Stewart Living Magazine	Martha Stewart's Encyclopedia of Crafts	Encyclopedia	
9	Geoffrey I. Webb	Encyclopedia of Machine Learning and Data Mining	Encyclopedia	
10	Claude Lecouteux	Encyclopedia of Norse and Germanic Folklore, Mythology, and Magic	Encyclopedia	
11	Emiko Konomi	Beginning Japanese for Professionals: Book 1	Textbook	
12	Yiping Fang, Vivek Shandas, Eugenio Arriaga	Spatial Thinking in Planning Practice: An Introduction to GIS	Textbook	

13	Chris Allen	The Balance of Personality	Textbook	
14	William Stallings, Lawrie Brown	Computer Security Principles and Practice.pdf	Textbook	
15	Alyn G. McFarland, Teresa Quigg, Nora Henry	AQA A-Level Chemistry Textbook Year 1	Textbook	

	ISBN or ASIN	Title	Author
1	9780044453420	The Concise Encyclopedia of Western Philosophy and Philosophers	J. O. Urmson (Editor); Jonathan Ree (Editor)
2	9780028658346	Encyclopedia of Science, Technology, and Ethics	Carl Mitcham
3	9780195123067	International Encyclopedia of Dance	Selma Jeanne Cohen (Editor); George Dorris (Editor); Thomas F. Kelly (Contribution by)
4	9780826409157	Encyclopedia of Women in Religious Art	Diane Apostolos-Cappadona (Editor)
5	9780028658506	Encyclopedia of Genocide and Crimes Against Humanity	Dinah Shelton
6	9780816049196	Historical Atlas of the Crusades	Angus Konstam
7	9780415188265	The State of Germany Atlas	Bernhard Schafers
8	9780528877537	Rand McNally Goode's World Atlas	Rand McNally Staff
9	9780762105106	Illustrated Atlas of the World	Reader's Digest Editors
10	9780816052196	Historical Atlas of the Holy Lands	Karen Farrengrton
11	9780933146389	Basic ESL Textbook	C. Sesma
12	9780198775768	The Oxford Textbook of	Keith Blois

		Marketing	
13	9780748715510	A Textbook of Holistic Aromatherapy	W. E. Arnould-Taylor
14	9780972151146	Success Dynamics Textbook	Geno Stampora
15	9780444865892	Economics of Organization	J.D. Hess

C. The datasets appear to have many books that were first published between 1926 and 1963.

93. There is no metadata showing the first publication date in any of the datasets.

94. However, I performed searches to identify possible evidence of a number of books in each of the datasets published between 1926 and 1963. To conduct this analysis, I used the *Goodreads Best Books of the Decade: 1940s*³⁵ and *Goodreads best books of the Decade: 1920s*³⁶ lists. I confirmed that each books' listed publication date on Goodreads fell between 1926 and 1963, filtering out 13 books from the 1920s list that did not, then searched for each of the remaining first 25 books on this list in all four datasets using the search procedures described above.

95. Of the 37 books, all published between 1926 and 1963 (with 25 published in the 1940s and 12 published between 1926 and 1930), I found potential matches to 33 of them in Anthropic's [REDACTED] Dataset, all 37 in Anthropic's [REDACTED] Dataset, 31 in Anthropic's [REDACTED] Dataset, and all 37 in the [REDACTED] Dataset. The following tables list the results of these searches.

	Author	Title	Publication Date	[REDACTED]	[REDACTED]
1	George	1984	1949	[REDACTED]	[REDACTED]

³⁵ *Best Books of the Decade: 1940s*, goodreads.com, (last visited Apr. 15, 2025), available at https://www.goodreads.com/list/show/23.Best_Books_of_the_Decade_1940s.

³⁶ *Best Books of the Decade: 1920s*, goodreads.com, (last visited Apr. 15, 2025), available at https://www.goodreads.com/list/show/39.Best_Books_of_the_Decade_1920s.

	Orwell				
2	George Orwell	Animal Farm	1945		
3					
4	3 Anne Frank	The Diary of a Young Girl	1947		
5					
6	4 Antoine de Saint-Exupéry	The Little Prince	1943		
7					
8	5 Albert Camus	The Stranger	1942		
9					
10	6 Tennessee Williams	A Streetcar Named Desire	1947		
11					
12					
13	7 Ernest Hemingway	For Whom the Bell Tolls	1940		
14					
15	8 Astrid Lindgren	Pippi Longstocking	1945		
16					
17					
18	9 Arthur Miller	Death of a Salesman	1949		
19					
20					
21	10 Betty Smith	A Tree Grows in Brooklyn	1943		
22					
23					
24	11 Albert Camus	The Plague	1947		
25					
26	12 Margaret Wise Brown	Goodnight Moon	1947		
27					
28	13 Evelyn Waugh	Brideshead Revisited	1945		

14	C.S. Lewis	The Screwtape Letters	1942		
15	Carson McCullers	The Heart is a Lonely Hunter	1940		
16	John Steinbeck	Cannery Row	1945		
17	John Steinbeck	The Pearl	1947		
18	Dodie Smith	I Capture the Castle	1948		
19	Alan Paton	Cry, the Beloved Country	1948		
20	John Hersey	Hiroshima	1946		
21	Ayn Rand	The Fountainhead	1943		
22	Jorge Luis Borges	Ficciones	1944		
23	Tennessee Williams	The Glass Menagerie	1945		
24	Robert McCloskey	Make Way for Ducklings	1941		
25	C.S. Lewis	The Chronicles of Narnia	1956		

1					
2	26	A.A. Milne	Winnie-the-Pooh	1926	
3					
4	27	Erich Maria Remarque	All Quiet on the Western Front	1928	
5					
6	28	Ernest Hemingway	The Sun Also Rises	1926	
7					
8	29	William Faulkner	The Sound and the Fury	1929	
9					
10					
11	30	Ernest Hemingway	A Farewell to Arms	1929	
12					
13					
14					
15	31	Virginia Woolf	To the Lighthouse	1927	
16					
17					
18	32	Hermann Hesse	Steppenwolf	1927	
19					
20	33	Agatha Christie	The Murder of Roger Ackroyd	1926	
21					
22	34	Virginia Woolf	A Room of One's Own	1929	
23					
24					
25	35	D.H. Lawrence	Lady Chatterley's Lover	1928	
26					
27	36	Virginia Woolf	Orlando	1928	
28					

37	Willa Cather	Death Comes for the Archbishop	1927		
----	--------------	--------------------------------	------	--	--

	Author	Title	Publication Date		
1	George Orwell	1984	1949		
2	George Orwell	Animal Farm	1945		
3	Anne Frank	The Diary of a Young Girl	1947		
4	Antoine de Saint-Exupéry	The Little Prince	1943		
5	Albert Camus	The Stranger	1942		
6	Tennessee Williams	A Streetcar Named Desire	1947		
7	Ernest Hemingway	For Whom the Bell Tolls	1940		
8	Astrid Lindgren	Pippi Longstocking	1945		
9	Arthur Miller	Death of a Salesman	1949		
10	Betty Smith	A Tree Grows in Brooklyn	1943		
11	Albert Camus	The Plague	1947		
12	Margaret Wise Brown	Goodnight Moon	1947		
13	Evelyn	Brideshead	1945		

	Waugh	Revisited			
14	C.S. Lewis	The Screwtape Letters	1942		
15	Carson McCullers	The Heart is a Lonely Hunter	1940		
16	John Steinbeck	Cannery Row	1945		
17	John Steinbeck	The Pearl	1947		
18	Dodie Smith	I Capture the Castle	1948		
19	Alan Paton	Cry, the Beloved Country	1948		
20	John Hersey	Hiroshima	1946		
21	Ayn Rand	The Fountainhead	1943		
22	Jorge Luis Borges	Ficciones	1944		
23	Tennessee Williams	The Glass Menagerie	1945		
24	Robert McCloskey	Make Way for Ducklings	1941		
25	C.S. Lewis	The Chronicles of Narnia	1956		
26	A.A. Milne	Winnie-the-Pooh	1926		
27	Erich Maria Remarque	All Quiet on the Western Front	1928		
28	Ernest Hemingway	The Sun Also Rises	1926		

29	William Faulkner	The Sound and the Fury	1929		
30	Ernest Hemingway	A Farewell to Arms	1929		
31	Virginia Woolf	To the Lighthouse	1927		
32	Hermann Hesse	Steppenwolf	1927		
33	Agatha Christie	The Murder of Roger Ackroyd	1926		
34	Virginia Woolf	A Room of One's Own	1929		
35	D.H. Lawrence	Lady Chatterley's Lover	1928		
36	Virginia Woolf	Orlando	1928		
37	Willa Cather	Death Comes for the Archbishop	1927		

	Author	Title	Publication Date	
1	George Orwell	1984	1949	
2	George Orwell	Animal Farm	1945	
3	Anne Frank	The Diary of a Young Girl	1947	
4	Antoine de Saint-Exupéry	The Little Prince	1943	
5	Albert Camus	The Stranger	1942	

6	Tennessee Williams	A Streetcar Named Desire	1947	
7	Ernest Hemingway	For Whom the Bell Tolls	1940	
8	Astrid Lindgren	Pippi Longstocking	1945	
9	Arthur Miller	Death of a Salesman	1949	
10	Betty Smith	A Tree Grows in Brooklyn	1943	
11	Albert Camus	The Plague	1947	
12	Margaret Wise Brown	Goodnight Moon	1947	
13	Evelyn Waugh	Brideshead Revisited	1945	
14	C.S. Lewis	The Screwtape Letters	1942	
15	Carson McCullers	The Heart is a Lonely Hunter	1940	
16	John Steinbeck	Cannery Row	1945	
17	John Steinbeck	The Pearl	1947	
18	Dodie Smith	I Capture the Castle	1948	
19	Alan Paton	Cry, the Beloved Country	1948	
20	John Hersey	Hiroshima	1946	
21	Ayn Rand	The Fountainhead	1943	
22	Jorge Luis	Ficciones	1944	

	Borges			
23	Tennessee Williams	The Glass Menagerie	1945	
24	Robert McCloskey	Make Way for Ducklings	1941	
25	C.S. Lewis	The Chronicles of Narnia	1956	
26	A.A. Milne	Winnie-the-Pooh	1926	
27	Erich Maria Remarque	All Quiet on the Western Front	1928	
28	Ernest Hemingway	The Sun Also Rises	1926	
29	William Faulkner	The Sound and the Fury	1929	
30	Ernest Hemingway	A Farewell to Arms	1929	
31	Virginia Woolf	To the Lighthouse	1927	
32	Hermann Hesse	Steppenwolf	1927	
33	Agatha Christie	The Murder of Roger Ackroyd	1926	
34	Virginia Woolf	A Room of One's Own	1929	
35	D.H. Lawrence	Lady Chatterley's Lover	1928	
36	Virginia Woolf	Orlando	1928	
37	Willa Cather	Death Comes for the Archbishop	1927	

	Author	Title	Publication Date	ISBN in
1	George Orwell	1984	1949	9780451524935
2	George Orwell	Animal Farm	1945	9780140008388
3	Anne Frank	The Diary of a Young Girl	1947	9780553296983
4	Antoine de Saint-Exupéry	The Little Prince	1943	9780590129275
5	Albert Camus	The Stranger	1942	9780394700021
6	Tennessee Williams	A Streetcar Named Desire	1947	9780451121806
7	Ernest Hemingway	For Whom the Bell Tolls	1940	9781856191524
8	Astrid Lindgren	Pippi Longstocking	1945	9780590016551
9	Arthur Miller	Death of a Salesman	1949	9780881030174
10	Betty Smith	A Tree Grows in Brooklyn	1943	9780060801267
11	Albert Camus	The Plague	1947	9780822010395
12	Margaret Wise Brown	Goodnight Moon	1947	9780060207052
13	Evelyn Waugh	Brideshead Revisited	1945	9780679423003
14	C.S. Lewis	The Screwtape Letters	1942	9780020867401
15	Carson McCullers	The Heart is a Lonely Hunter	1940	9780395079782
16	John Steinbeck	Cannery Row	1943	9780330241458
17	John	The Pearl	1945	9780670545759

	Steinbeck			
18	Dodie Smith	I Capture the Castle	1948	9781860491023
19	Alan Paton	Cry, the Beloved Country	1948	9780684718637
20	John Hersey	Hiroshima	1946	9780394548449
21	Ayn Rand	The Fountainhead	1943	9780451175120
22	Jorge Luis Borges	Ficciones	1944	9780679422990
23	Tennessee Williams	The Glass Menagerie	1945	9780451166364
24	Robert McCloskey	Make Way for Ducklings	1941	9780670451494
25	C.S. Lewis	The Chronicles of Narnia	1956	9780590254793
26	A.A. Milne	Winnie-the-Pooh	1926	9780140361216
27	Erich Maria Remarque	All Quiet on the Western Front	1928	9780316739924
28	Ernest Hemingway	The Sun Also Rises	1926	9781555460457
29	William Faulkner	The Sound and the Fury	1929	9780394700052
30	Ernest Hemingway	A Farewell to Arms	1929	9780586044711
31	Virginia Woolf	To the Lighthouse	1927	9780586044506
32	Hermann Hesse	Steppenwolf	1927	9780030121357
33	Agatha Christie	The Murder of Roger Ackroyd	1926	9780553350012
34	Virginia Woolf	A Room of One's Own	1929	9780156787338
35	D.H.	Lady Chatterley's	1928	9780451522474

	Lawrence	Lover		
36	Virginia Woolf	Orlando	1928	9780932986399
37	Willa Cather	Death Comes for the Archbishop	1927	9780896215061

96. These results were found by a targeted search for evidence of particular books published between 1926 and 1963, and do not represent a complete accounting of possible evidence of books published in this period in the book datasets. Nonetheless, I believe it is reasonable to conclude that there are many possible books in all four datasets that were published between 1926 and 1963.

D. The datasets have many books that were first published between 1963 and 1989.

97. I also performed searches to identify possible evidence of a number of books in each of the four datasets published between 1963 and 1989. To conduct this analysis, I used the *Goodreads Best Books of the Decade: 1970s* list.³⁷ I searched for each of the first 25 books on this list in all four datasets using the search procedures described above.

98. Of the 25 books, all published between 1963 and 1989, I found potential matches to 15 of them in Anthropic's [REDACTED] dataset, 22 in Anthropic's [REDACTED] Dataset, 21 in Anthropic's [REDACTED] Dataset, and all 25 in Anthropic's [REDACTED] Dataset, as follows:

	Author	Title	Publication Date	[REDACTED]	[REDACTED]
1	Douglas Adams	The Hitchhiker's Guide to the Galaxy	1979	[REDACTED]	[REDACTED]
2	Richard Adams	Watership Down	1972	[REDACTED]	[REDACTED]

³⁷ *Best Books of the Decade: 1970s*, goodreads.com, (last visited Apr. 15, 2025), available at <https://www.goodreads.com/list/show/21>.

3	Stephen King	The Stand	1978		
4	Stephen King	The Shining	1977		
5	William Goldman	The Princess Bride	1973		
6	Anne Rice	Interview with the Vampire	1976		
7	Shel Silverstein	Where the Sidewalk Ends	1974		
8	John Irving	The World According to Garp	1978		
9	Colleen McCullough	The Thorn Birds	1977		
10	Stephen King	Carrie	1974		
11	James Clavell	Shogun	1975		
12	Alex Haley	Roots: The Saga of an American Family	1976		
13	Stephen King	Salem's Lot	1975		
14	Dr. Seuss	The Lorax	1971		
15	Hunter S. Thompson	Fear and Loathing in Las Vegas: A Savage Journey to the Heart of the American Dream	1971		

16	Katherine Paterson	Bridge to Terabithia	1977	[REDACTED]	[REDACTED]
17	James Herriot	All Creatures Great and Small	1972	[REDACTED]	[REDACTED]
18	Michael Ende	The Neverending Story	1979	[REDACTED]	[REDACTED]
19	J.R.R. Tolkien	The Silmarillion	1977	[REDACTED]	[REDACTED]
20	William Peter Blatty	The Exorcist	1971	[REDACTED]	[REDACTED]
21	William Styron	Sophie's Choice	1979	[REDACTED]	[REDACTED]
22	Kurt Vonnegut	Breakfast of Champions	1973	[REDACTED]	[REDACTED]
23	John Le Carré	Tinker, Tailor, Soldier, Spy	1974	[REDACTED]	[REDACTED]
24	Dee Brown	Bury My Heart at Wounded Knee: An Indian History of the American West	1970	[REDACTED]	[REDACTED]
25	Beatrice Sparks	Go Ask Alice	1971	[REDACTED]	[REDACTED]

[REDACTED]

	Author	Title	Publication Date	[REDACTED]	[REDACTED]
1	Douglas Adams	The Hitchhiker's Guide to the Galaxy	1979	Fiction	[REDACTED]
2	Richard Adams	Watership Down	1972	Fiction	[REDACTED]
3	Stephen King	The Stand	1978	Fiction	[REDACTED]

4	Stephen King	The Shining	1977	Fiction	
5	William Goldman	The Princess Bride	1973	Fiction	
6	Anne Rice	Interview with the Vampire	1976	Fiction	
7	Shel Silverstein	Where the Sidewalk Ends	1974	Fiction	
8	John Irving	The World According to Garp	1978	Fiction	
9	Colleen McCullough	The Thorn Birds	1977	Fiction	
10	Stephen King	Carrie	1974	Fiction	
11	James Clavell	Shogun	1975	Fiction	
12	Alex Haley	Roots: The Saga of an American Family	1976	Fiction	
13	Stephen King	Salem's Lot	1975	Fiction	
14	Dr. Seuss	The Lorax	1971	Fiction	
15	Hunter S. Thompson	Fear and Loathing in Las Vegas: A Savage Journey to the Heart of the American Dream	1971	Fiction	
16	Katherine Paterson	Bridge to Terabithia	1977	Fiction	
17	James Herriot	All Creatures Great and Small	1972	Fiction	
18	Michael Ende	The Neverending Story	1979	Fiction	

19	J.R.R. Tolkein	The Silmarillion	1977	Fiction	[REDACTED]
20	William Peter Blatty	The Exorcist	1971	Fiction	[REDACTED]
21	William Styron	Sophie's Choice	1979	Fiction	[REDACTED]
22	Kurt Vonnegut	Breakfast of Champions	1973	Fiction	[REDACTED]
23	John Le Carré	Tinker, Tailor, Soldier, Spy	1974	Fiction	[REDACTED]
24	Dee Brown	Bury My Heart at Wounded Knee: An Indian History of the American West	1970	Fiction	[REDACTED]
25	Beatrice Sparks	Go Ask Alice	1971	Fiction	[REDACTED]

	Author	Title	Publication Date	[REDACTED]
1	Douglas Adams	The Hitchhiker's Guide to the Galaxy	1979	[REDACTED]
2	Richard Adams	Watership Down	1972	[REDACTED]
3	Stephen King	The Stand	1978	[REDACTED]
4	Stephen King	The Shining	1977	[REDACTED]
5	William Goldman	The Princess Bride	1973	[REDACTED]
6	Anne Rice	Interview with the Vampire	1976	[REDACTED]

7	Shel Silverstein	Where the Sidewalk Ends	1974	
8	John Irving	The World According to Garp	1978	
9	Colleen McCullough	The Thorn Birds	1977	
10	Stephen King	Carrie	1974	
11	James Clavell	Shogun	1975	
12	Alex Haley	Roots: The Saga of an American Family	1976	
13	Stephen King	Salem's Lot	1975	
14	Dr. Seuss	The Lorax	1971	
15	Hunter S. Thompson	Fear and Loathing in Las Vegas: A Savage Journey to the Heart of the American Dream	1971	
16	Katherine Paterson	Bridge to Terabithia	1977	
17	James Herriot	All Creatures Great and Small	1972	
18	Michael Ende	The Neverending Story	1979	
19	J.R.R. Tolkien	The Silmarillion	1977	
20	William Peter Blatty	The Exorcist	1971	
21	William Styron	Sophie's Choice	1979	
22	Kurt	Breakfast of	1973	

	Vonnegut	Champions		
23	John Le Carré	Tinker, Tailor, Soldier, Spy	1974	
24	Dee Brown	Bury My Heart at Wounded Knee: An Indian History of the American West	1970	
25	Beatrice Sparks	Go Ask Alice	1971	

	Author	Title	Publication Date	ISBN in
1	Douglas Adams	The Hitchhiker's Guide to the Galaxy	1979	9780330258647
2	Richard Adams	Watership Down	1972	9780380395866
3	Stephen King	The Stand	1978	9780451139719
4	Stephen King	The Shining	1977	9780451160911
5	William Goldman	The Princess Bride	1973	9780345301147
6	Anne Rice	Interview with the Vampire	1976	9780345337665
7	Shel Silverstein	Where the Sidewalk Ends	1974	9780060256685
8	John Irving	The World According to Garp	1978	9780525237709
9	Colleen McCullough	The Thorn Birds	1977	9780060129569
10	Stephen King	Carrie	1974	9781856953184
11	James Clavell	Shogun	1975	9780440178002
12	Alex Haley	Roots: The Saga of	1976	9780385037877

		an American Family		
13	Stephen King	Salem's Lot	1975	9780451168085
14	Dr. Seuss	The Lorax	1971	9780394823379
15	Hunter S. Thompson	Fear and Loathing in Las Vegas: A Savage Journey to the Heart of the American Dream	1971	9780679724193
16	Katherine Paterson	Bridge to Terabithia	1977	9780380432813
17	James Herriot	All Creatures Great and Small	1972	9780553204346
18	Michael Ende	The Neverending Story	1979	9780140074314
19	J.R.R. Tolkein	The Silmarillion	1977	9780345272553
20	William Peter Blatty	The Exorcist	1971	9780061007224
21	William Styron	Sophie's Choice	1979	9780394461090
22	Kurt Vonnegut	Breakfast of Champions	1973	9780440131489
23	John Le Carré	Tinker, Tailor, Soldier, Spy	1974	9781444705652
24	Dee Brown	Bury My Heart at Wounded Knee: An Indian History of the American West	1970	9780330245692
25	Beatrice Sparks	Go Ask Alice	1971	9780380005239

99. These results were found by a targeted search for evidence of particular books published between 1963 and 1989, and do not represent a complete accounting of possible evidence of books published in this period in the book datasets. Nonetheless, I believe it is

reasonable to conclude that there are many possible books in all four datasets that were published between 1963 and 1989.

E. The datasets appear to contain many books likely written by foreign authors.

100. There is no metadata showing the nationality of the authors in any of the Internet Books Datasets.

101. However, I performed searches to identify possible evidence of a number of books in each of the four datasets written by foreign authors. I used the *Goodreads Non American Books that Every American Should Read*³⁸ list for this search. I searched for each of the first 20 books on this list written by authors who I believe, based on publicly-available information, lived into the 2000s and are foreign authors in all four datasets using the search procedures described above. I additionally looked for corroborating publicly-available information about each author's nationality by searching for the author's Wikipedia page and noting all listed nationalities.³⁹

102. Of the 20 books, all written by individuals, whose authors may be nationals of countries other than the United States, I found 17 of them in [REDACTED], 19 in [REDACTED], 18 in [REDACTED], and 19 in the [REDACTED], as follows:

	Title	Author	Nationality	[REDACTED]	[REDACTED]
1	Disgrace	J.M. Coetzee	South Africa	[REDACTED]	[REDACTED]
2	A Long Way Gone: Memoirs of a	Ishmael Beah	Sierra Leone	[REDACTED]	[REDACTED]

³⁸ *Non American Books that Every American Should Read*, goodreads.com (last visited Apr. 15, 2025), available at https://www.goodreads.com/list/show/5534.Non_American_books_that_every_American_should_read.

³⁹ One author from the Goodreads list, Ashok Kallarakkal, does not have a Wikipedia page. His Amazon author biography page notes that he is a resident of India.

⁴⁰ For readability, I omit the ".epub.txt" that concludes each file name in the provided [REDACTED] dataset from these tables.

	Boy Soldier				
3	The God of Small Things	Arundhati Roy	India		
4	Things Fall Apart	Chinua Achebe	Nigeria		
5	Half of a Yellow Sun	Chimamanda Ngozi Adicheie	Nigeria		
6	Persepolis	Marjane Satrapi	Iran, France		
7	A Fine Balance	Rohinton Mistry	Canada		
8	Long Walk to Freedom	Nelson Mandela	South Africa		
9	Atonement	Ian McEwan	UK		
10	The Tin Drum	Gunter Grass	Germany		
11	Life of Pi	Yann Martel	Canada		
12	The No. 1 Ladies' Detective Agency	Alexander McCall Smith	UK		
13	Little Bee	Chris Cleave	UK		
14	The Name of the Rose	Umberto Eco	Italy		
15	The Book Thief	Markus Zusak	Germany, Australia		

16	The Girl With the Dragon Tattoo	Steig Larsson	Sweden		
17	The God Delusion	Richard Dawkins	UK		
18	Norwegian Wood	Haruki Murakami	Japan		
19	The Shadow of the Wind	Carlos Ruiz Zafon	Spain		
20	Curiosity kills the KatHA	Ashok Kallarakkal	India		

	Title	Author	Nationality		
1	Disgrace	J.M. Coetzee	South Africa		
2	A Long Way Gone: Memoirs of a Boy Soldier	Ishmael Beah	Sierra Leone		
3	The God of Small Things	Arundhati Roy	India		
4	Things Fall Apart	Chinua Achebe	Nigeria		
5	Half of a Yellow Sun	Chimamanda Ngozi Adicheie	Nigeria		
6	Persepolis	Marjane	Iran, France		

		Satrapi			
7	A Fine Balance	Rohinton Mistry	Canada		
8	Long Walk to Freedom	Nelson Mandela	South Africa		
9	Atonement	Ian McEwan	UK		
10	The Tin Drum	Gunter Grass	Germany		
11	Life of Pi	Yann Martel	Canada		
12	The No. 1 Ladies' Detective Agency	Alexander McCall Smith	UK		
13	Little Bee	Chris Cleave	UK		
14	The Name of the Rose	Umberto Eco	Italy		
15	The Book Thief	Markus Zusak	Germany, Australia		
16	The Girl With the Dragon Tattoo	Steig Larsson	Sweden		
17	The God Delusion	Richard Dawkins	UK		
18	Norwegian Wood	Haruki Murakami	Japan		
19	The Shadow of the Wind	Carlos Ruiz Zafon	Spain		
20	Curiosity kills the KatHA	Ashok Kallarakkal	India		

	Title	Author	Nationality	
1	Disgrace	J.M. Coetzee	South Africa	
2	A Long Way Gone: Memoirs of a Boy Soldier	Ishmael Beah	Sierra Leone	
3	The God of Small Things	Arundhati Roy	India	
4	Things Fall Apart	Chinua Achebe	Nigeria	
5	Half of a Yellow Sun	Chimamanda Ngozi Adichie	Nigeria	
6	Persepolis	Marjane Satrapi	Iran, France	
7	A Fine Balance	Rohinton Mistry	Canada	
8	Long Walk to Freedom	Nelson Mandela	South Africa	
9	Atonement	Ian McEwan	UK	
10	The Tin Drum	Gunter Grass	Germany	
11	Life of Pi	Yann Martel	Canada	
12	The No. 1 Ladies' Detective Agency	Alexander McCall Smith	UK	
13	Little Bee	Chris Cleave	UK	
14	The Name of the Rose	Umberto Eco	Italy	
15	The Book	Markus Zusak	Germany,	

	Thief		Australia	
16	The Girl With the Dragon Tattoo	Steig Larsson	Sweden	
17	The God Delusion	Richard Dawkins	UK	
18	Norwegian Wood	Haruki Murakami	Japan	
19	The Shadow of the Wind	Carlos Ruiz Zafon	Spain	
20	Curiosity kills the KatHA	Ashok Kallarakkal	India	

	Title	Author	Nationality	ISBN in Metadata
1	Disgrace	J.M. Coetzee	South Africa	9780670887316
2	A Long Way Gone: Memoirs of a Boy Soldier	Ishmael Beah	Sierra Leone	9780374105235
3	The God of Small Things	Arundhati Roy	India	9780783882963
4	Things Fall Apart	Chinua Achebe	Nigeria	9780435900014
5	Half of a Yellow Sun	Chimamanda Ngozi Adicheie	Nigeria	9781400044160
6	Persepolis	Marjane Satrapi	Iran, France	9780375422300
7	A Fine Balance	Rohinton Mistry	Canada	9780771060526
8	Long Walk to Freedom	Nelson Mandela	South Africa	9780316545853
9	Atonement	Ian McEwan	UK	9780887846410
10	The Tin Drum	Gunter Grass	Germany	9780394703008

11	Life of Pi	Yann Martel	Canada	9780547848419
12	The No. 1 Ladies' Detective Agency	Alexander McCall Smith	UK	9780748662524
13	Little Bee	Chris Cleave	UK	9781416589648
14	The Name of the Rose	Umberto Eco	Italy	9780151446476
15	The Book Thief	Markus Zusak	Germany, Australia	9780375831003
16	The Girl With the Dragon Tattoo	Steig Larsson	Sweden	9781847246929
17	The God Delusion	Richard Dawkins	UK	9780593055489
18	Norwegian Wood	Haruki Murakami	Japan	9780375704024
19	The Shadow of the Wind	Carlos Ruiz Zafon	Spain	9780752859545
20	Curiosity kills the KatHA	Ashok Kallarakkal	India	Not Found

I declare under penalty of perjury under the laws of the United States of America that the foregoing is true and correct. Executed on April 17, 2025 in Clarksville, Maryland.

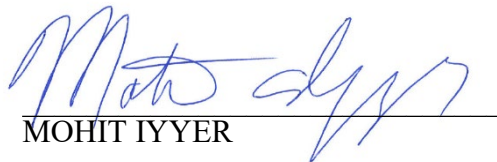

MOHIT IYER

EXHIBIT A**MOHIT IYER**miyyer@cs.umass.edu
<http://cs.umass.edu/~miyyer>ACADEMIC APPOINTMENTS

University of Massachusetts, Amherst

Associate Professor, Computer Science

2023 — *present*

Assistant Professor, Computer Science

2018 — 2023

Research Interests: *natural language processing, machine learning*EDUCATION

University of Maryland, College Park

2012 — 2017

Ph.D. in Computer Science

Master of Science, Computer Science, 2012 - 2014

Washington University in St. Louis

2008 — 2012

BS in Computer Science

Minors: Bioinformatics & Creative Writing

RESEARCH & PROFESSIONAL EXPERIENCE

Postdoctoral Researcher, [Allen Institute for AI](#)

2017 — 2018

Supervisors: [Luke Zettlemoyer](#), [Scott Yih](#)**Research Assistant**, [University of Maryland, College Park](#)

2013 — 2017

Advisors: [Jordan Boyd-Graber](#), [Hal Daumé III](#)**Research Intern**, [NLP Group, Microsoft Research, Redmond](#)

Summer 2016

Supervisors: [Scott Yih](#), [Ming-Wei Chang](#)**Research Intern**, [MetaMind](#)

Spring 2015

Supervisor: [Richard Socher](#)AWARDS

Distinguished Paper Award, ACM CCS 2023

Outstanding Paper Award, EACL 2023

Samsung AI Researcher of the Year Award, 2022

NSF CAREER Award, 2021

Best Long Paper Award, NAACL 2018

Larry S. Davis Doctoral Dissertation Award, 2017

Best Long Paper Award, NAACL 2016

Best Demonstration Award, NeurIPS 2015

CONFERENCE PUBLICATIONS

As of December 2024, my papers have been cited over 27K times (h-index 42) according to my [Google Scholar profile](#). ACL, NAACL, EMNLP, ICML, NeurIPS, CVPR, and ICLR are peer reviewed conferences with acceptance rates typically around 25-30%.

1. Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, **Mohit Iyyer**. [One Thousand and One Pairs: A "novel" challenge for long-context language models](#). *EMNLP 2024*.
2. Yapei Chang, Kalpesh Krishna, Amir Houmansadr, John Wieting, **Mohit Iyyer**. [PostMark: A Robust Blackbox Watermark for Large Language Models](#). *EMNLP 2024*.
3. Yixiao Song, Yekyung Kim, **Mohit Iyyer**. [VERISCORE: Evaluating the factuality of verifiable claims in long-form text generation](#). *Findings of EMNLP 2024*.
4. Chau Minh Pham, Simeng Sun, **Mohit Iyyer**. [Suri: Multi-constraint Instruction Following for Long-form Text Generation](#). *Findings of EMNLP 2024*.
5. Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, **Mohit Iyyer**. [FABLES: Evaluating faithfulness and content selection in book-length summarization](#). *COLM 2024*.
6. Ali Naseh, Katherine Thai, **Mohit Iyyer**, Amir Houmansadr. [Iteratively Prompting Multimodal LLMs to Reproduce Natural and AI-Generated Images](#). *COLM 2024*.
7. Jiachen Zhao, Wenlong Zhao, Andrew Drozdov, Benjamin Rozonoyer, Md Arafat Sultan, Jay-Yoon Lee, **Mohit Iyyer**, Andrew McCallum. [Multistage Collaborative Knowledge Distillation from a Large Language Model for Semi-Supervised Sequence Generation](#). *ACL 2024*.
8. Tu Vu, **Mohit Iyyer**, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, Thang Luong. [FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation](#). *Findings of ACL 2024*.
9. Chau Pham, Alexander Hoyle, Simeng Sun, **Mohit Iyyer**. [TopicGPT: A Prompt-based Topic Modeling Framework](#). *NAACL 2024*.
10. Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, Kevin Gimpel, **Mohit Iyyer**. [GEE! Grammar Error Explanation with Large Language Models](#). *Findings of NAACL 2024*.
11. Prasanna Lakkur Subramanyam, **Mohit Iyyer**, Brian Levine. [Triage of Messages and Conversations in a Large-Scale Child Victimization Corpus](#). *ACM The Web Conference 2024 (Web4Good Track)*.
12. Yapei Chang, Kyle Lo, Tanya Goyal, **Mohit Iyyer**. [BoooookScore: A systematic exploration of book-length summarization in the era of LLMs](#). *ICLR 2024 (oral)*.
13. Simeng Sun, Yang Liu, Shuohang Wang, Chenguang Zhu, **Mohit Iyyer**. [PEARL: Prompting Large Language Models to Plan and Execute Actions Over Long Documents](#). *EACL 2024*.
14. Simeng Sun, Yang Liu, Dan Iter, Chenguang Zhu, **Mohit Iyyer**. [How Does In-Context Learning Help Prompt Tuning?](#) *Findings of EACL 2024*.

15. Shufan Wang, Yixiao Song, Andrew Drozdov, Aparna Garimella, Varun Manjunatha, **Mohit Iyyer**. [kNN-LM Does Not Improve Open-ended Text Generation](#). *EMNLP 2023*.
16. Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, **Mohit Iyyer**, Luke Zettlemoyer, Hannaneh Hajishirzi. [FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation](#). *EMNLP 2023*.
17. Nader Akoury, Qian Yang, **Mohit Iyyer**. [Disco Elysium: Exploring Player Perceptions of LLM-Generated Dialogue within a Commercial Video Game](#). *Findings of EMNLP 2023*.
18. Andrew Drozdov, Honglei Zhuang, Zhuyun Dai, Zhen Qin, Razieh Rahimi, Xuanhui Wang, Dana Alon, **Mohit Iyyer**, Andrew McCallum, Donald Metzler, Kai Hui. [PaRaDe: Passage Ranking using Demonstrations with LLMs](#). *Findings of EMNLP 2023 (short)*.
19. Marzena Karpinska, **Mohit Iyyer**. [Large language models effectively leverage document-level context for literary translation, but critical errors persist](#). *WMT 2023*.
20. Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, **Mohit Iyyer**. [Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense](#). *NeurIPS 2023*.
21. Fangyuan Xu*, Yixiao Song*, **Mohit Iyyer**, Eunsol Choi. [A Critical Evaluation of Evaluations for Long-form Question Answering](#). *ACL 2023*.
22. Ali Naseh, Kalpesh Krishna, **Mohit Iyyer**, Amir Houmansadr. [Stealing the Decoding Algorithms of Language Models](#). *CCS 2023*. *Distinguished paper award*.
23. Kalpesh Krishna, Erin Bransom, Bailey Kuehl, **Mohit Iyyer**, Pradeep Dasigi, Arman Cohan, Kyle Lo. [LongEval: Guidelines for Human Evaluation of Faithfulness in Long-form Summarization](#). *EACL 2023*. *Outstanding paper award*.
24. Ankita Gupta, Marzena Karpinska, Wenlong Zhao, Kalpesh Krishna, Jack Merullo, Luke Yeh, **Mohit Iyyer**, Brendan O'Connor. [ezCoref: Towards Unifying Annotation Guidelines for Coreference Resolution](#). *Findings of EACL 2023*.
25. Kalpesh Krishna, Yapei Chang, John Wieting, **Mohit Iyyer**. [RankGen: Improving Text Generation with Large Ranking Models](#). *EMNLP 2022*.
26. Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, **Mohit Iyyer**, Noah Constant. [Overcoming Catastrophic Forgetting in Zero-Shot Cross-Lingual Generation](#). *EMNLP 2022*.
27. Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, **Mohit Iyyer**. [SLING: Sino Linguistic Evaluation of Large Language Models](#). *EMNLP 2022*.
28. Katherine Thai*, Marzena Karpinska*, Kalpesh Krishna, William Ray, Moira Inghilleri, John Wieting, **Mohit Iyyer**. [Exploring Document-Level Literary Machine Translation with Parallel Paragraphs from World Literature](#). *EMNLP 2022*.
29. Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, **Mohit Iyyer**. [DEMETER: Diagnosing Evaluation Metrics for Translation](#). *EMNLP 2022*.
30. Andrew Drozdov, Shufan Wang, Razieh Rahimi, Andrew McCallum, Hamed Zamani, **Mohit Iyyer**. [You can't pick your neighbors, or can you? When and How to Rely on Retrieval in the KNN-LM](#). *Findings of EMNLP 2022*.

31. Shufan Wang, Fangyuan Xu, Laure Thompson, Eunsol Choi, **Mohit Iyyer**. [Modeling Exemplification in Long-form Question Answering via Retrieval](#). *NAACL* 2022.
32. Simeng Sun, Katherine Thai, **Mohit Iyyer**. [ChapterBreak: A Challenge Dataset for Long-Range Language Models](#). *NAACL* 2022 (short).
33. Katherine Thai, Yapei Chang, Kalpesh Krishna, **Mohit Iyyer**. [RELIC: Retrieving Evidence for Literary Claims](#). *ACL* 2022.
34. Simeng Sun, Kalpesh Krishna, Andrew Mattarella-Micke, and **Mohit Iyyer**. [Do Long-Range Language Models Actually Use Long-Range Context?](#) *EMNLP* 2021.
35. Tu Vu, Minh-Thang Luong, Quoc Le, Grady Simon, and **Mohit Iyyer**. [STraTA: Self-Training with Task Augmentation for Better Few-shot Learning](#). *EMNLP* 2021.
36. Marzena Karpinska, Nader Akoury, and **Mohit Iyyer**. [The Perils of Using Mechanical Turk to Evaluate Open-Ended Text Generation](#). *EMNLP* 2021.
37. Shufan Wang, Laure Thompson, and **Mohit Iyyer**. [Phrase-BERT: Improved Phrase Embeddings from BERT with an Application to Corpus Exploration](#). *EMNLP* 2021.
38. Zhiyang Xu, Andrew Drozdov, Jay Yoon Lee, Tim O’Gorman, Subendhu Rongali, Dylan Finkbeiner, Shilpa Suresh, **Mohit Iyyer**, and Andrew McCallum. [Improved Latent Tree Induction with Distant Supervision via Span Constraints](#). *EMNLP* 2021.
39. Simeng Sun, Wenlong Zhao, Varun Manjunatha, Rajiv Jain, Vlad Morariu, Franck Dernoncourt, Balaji Vasan Srinivasan, and **Mohit Iyyer**. [IGA: An Intent-Guided Authoring Assistant](#). *EMNLP* 2021.
40. Sumanta Bhattacharyya, Amirmohammad Rooshenas, Subhajit Naskar, Simeng Sun, **Mohit Iyyer**, Andrew McCallum. [Energy-Based Reranking: Improving Neural Machine Translation Using Energy-Based Models](#). *ACL* 2021.
41. Iman Deznabi, **Mohit Iyyer**, Madalina Fiterau. [Predicting In-Hospital Mortality by Combining Clinical Notes with Time-Series Data](#). *Findings of ACL* 2021.
42. Ameet Trivedi, Kate Silverstein, Emma Strubell, **Mohit Iyyer**, Prashant Shenoy. [WiFiMod: Transformer-based Indoor Human Mobility Modeling using Passive Sensing](#). *ACM COMPASS* 2021.
43. Kalpesh Krishna, Aurko Roy, **Mohit Iyyer**. [Hurdles to Progress in Long-form Question Answering](#). *NAACL* 2021.
44. Hiroshi Iida, June Thai, Varun Manjunatha, **Mohit Iyyer**. [TABBIE: Pretrained Representations of Tabular Data](#). *NAACL* 2021.
45. Simeng Sun, **Mohit Iyyer**. [Revisiting Simple Neural Probabilistic Language Models](#). *NAACL* 2021 (short).
46. Haw-Shiuan Chang, Jiaming Yuan, **Mohit Iyyer**, Andrew McCallum. [Changing the Mind of Transformers for Topically-Controllable Language Generation](#). *EACL* 2021.
47. Chen Qu, Liu Yang, Cen Chen, W. Bruce Croft, Kalpesh Krishna, **Mohit Iyyer**. [Weakly-Supervised Open-Retrieval Conversational Question Answering](#). *ECIR* 2021.

48. Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, **Mohit Iyyer**. [Exploring and Predicting Transferability across NLP Tasks](#). *EMNLP* 2020.
49. Kalpesh Krishna, John Wieting, **Mohit Iyyer**. [Reformulating Unsupervised Style Transfer as Paraphrase Generation](#). *EMNLP* 2020.
50. Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, **Mohit Iyyer**. [STORIUM: A Dataset and Platform for Machine-in-the-Loop Story Generation](#). *EMNLP* 2020.
51. Andrew Drozdov, Subendhu Rongali, Yi-Pei Chen, Tim O’Gorman, **Mohit Iyyer**, Andrew McCallum. [Unsupervised Parsing with S-DIORA: Single Tree Encoding for Deep Inside-Outside Recursive Autoencoders](#). *EMNLP* 2020.
52. Weiqiu You*, Simeng Sun*, **Mohit Iyyer**. [Hard-Coded Gaussian Attention for Neural Machine Translation](#). *ACL* 2020.
53. Fenfei Guo, Jordan Boyd-Graber, **Mohit Iyyer**, Leah Findlater. [Which Evaluations Uncover Sense Representations that Actually Make Sense?](#) *LREC* 2020.
54. Kalpesh Krishna, Gaurav Singh Tomar, Ankur Parikh, Nicolas Papernot, **Mohit Iyyer**. [Thieves on Sesame Street! Model Extraction of BERT-based APIs](#). *ICLR* 2020.
55. Jack Merullo*, Luke Yeh*, Abram Handler, Alvin Grissom II, Brendan O’Connor, **Mohit Iyyer**. [Investigating Sports Commentator Bias within a Large Corpus of American Football Broadcasts](#). *EMNLP* 2019 (short).
56. Andrew Drozdov, Patrick Verga, Yi-Pei Chen, **Mohit Iyyer**, Andrew McCallum. [Unsupervised Labeled Parsing with Deep Inside-Outside Recursive Autoencoders](#). *EMNLP* 2019 (short).
57. Chen Qu, Liu Yang, Minghui Qiu, Yongfeng Zhang, Cen Chen, W. Bruce Croft, **Mohit Iyyer**. [Attentive History Selection for Conversational Question Answering](#). *CIKM* 2019.
58. Nader Akoury, Kalpesh Krishna, **Mohit Iyyer**. [Syntactically Supervised Transformers for Faster Neural Machine Translation](#). *ACL* 2019.
59. Kalpesh Krishna, **Mohit Iyyer**. [Generating Question-Answer Hierarchies](#). *ACL* 2019.
60. Tu Vu, **Mohit Iyyer**. [Encouraging Paragraph Embeddings to Remember Sentence Identity Improves Classification](#). *ACL* 2019 (short).
61. Chen Qu, Liu Yang, Minghui Qiu, W. Bruce Croft, Yongfeng Zhang, **Mohit Iyyer**. [BERT with History Modeling for Conversational Question Answering](#). *SIGIR* 2019 (short).
62. Andrew Drozdov, Patrick Verga, Mohit Yadav, **Mohit Iyyer**, Andrew McCallum. [Unsupervised Latent Tree Induction with Deep Inside-Outside Recursive Auto-Encoders](#). *NAACL* 2019.
63. Shufan Wang, **Mohit Iyyer**. [Casting Light on Invisible Cities: Computationally Engaging with Literary Criticism](#). *NAACL* 2019 (short).
64. Eunsol Choi*, He He*, **Mohit Iyyer***, Mark Yatskar*, Wen-tau Yih, Yejin Choi, Percy Liang, Luke Zettlemoyer. [QuAC: Question Answering in Context](#). *EMNLP* 2018.

65. Shi Feng, Eric Wallace, Alvin Grissom II, **Mohit Iyyer**, Pedro Rodriguez, Jordan Boyd-Graber. [Pathologies of Neural Models Make Interpretation Difficult](#). *EMNLP* 2018.
66. Kalpesh Krishna, Preethi Jyothi, **Mohit Iyyer**. [Revisiting the Importance of Encoding Logic Rules in Sentiment Classification](#). *EMNLP* 2018 (short).
67. **Mohit Iyyer***, John Wieting*, Kevin Gimpel, Luke Zettlemoyer. [Adversarial Example Generation with Syntactically Controlled Paraphrase Networks](#). *NAACL* 2018.
68. Matthew E. Peters, Mark Neumann, **Mohit Iyyer**, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer. [Deep contextualized word representations](#). *NAACL* 2018. *Best long paper award*
69. Varun Manjunatha*, **Mohit Iyyer***, Jordan Boyd-Graber, Larry Davis. [Learning to Color from Language](#). *NAACL* 2018 (short); also appeared at *NIPS 2017 ViGIL Workshop*.
70. **Mohit Iyyer**, Wen-tau Yih, Ming-Wei Chang. [Search-based Neural Structured Learning for Sequential Question Answering](#). *ACL* 2017.
71. **Mohit Iyyer***, Varun Manjunatha*, Anupam Guha, Yogarshi Vyas, Jordan Boyd-Graber, Hal Daumé III, Larry Davis. [The Amazing Mysteries of the Gutter: Drawing Inferences Between Panels in Comic Book Narratives](#). *CVPR* 2017.
72. Snigdha Chaturvedi, **Mohit Iyyer**, Hal Daumé III. [Unsupervised Learning of Evolving Relationships Between Literary Characters](#). *AAAI* 2017.
73. **Mohit Iyyer**, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, Hal Daumé III. [Feuding Families and Former Friends: Unsupervised Learning for Dynamic Fictional Relationships](#). *NAACL* 2016. *Best long paper award*.
74. Ankit Kumar, Ozan Irsoy, Peter Ondruska, **Mohit Iyyer**, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. [Ask Me Anything: Dynamic Memory Networks for Natural Language Processing](#). *ICML* 2016.
75. Jordan Boyd-Graber, **Mohit Iyyer**, He He, Hal Daumé III. Interactive Incremental Question Answering. *NIPS Demonstration Track*, 2015. *Outstanding Demonstration Award*.
76. **Mohit Iyyer**, Varun Manjunatha, Jordan Boyd-Graber, Hal Daumé III. [Deep Unordered Composition Rivals Syntactic Methods for Text Classification](#). *ACL* 2015.
77. Anupam Guha, **Mohit Iyyer**, Danny Bouman, Jordan Boyd-Graber. [Removing the Training Wheels: A Coreference Dataset that Entertains Humans and Challenges Computers](#). *NAACL* 2015.
78. **Mohit Iyyer**, Jordan Boyd-Graber, Richard Socher, Hal Daumé III. [A Neural Network for Factoid Question Answering over Paragraphs](#). *EMNLP* 2014.
79. **Mohit Iyyer**, Peter Enns, Jordan Boyd-Graber, Philip Resnik. [Political Ideology Detection Using Recursive Neural Networks](#). *ACL* 2014.

WORKSHOP PUBLICATIONS

Nader Akoury, Ronan Salz, **Mohit Iyyer**. [Towards Grounded Dialogue Generation in Video Game Environments](#) *AAAI Workshop on Creative AI Across Modalities*, 2023.

Simeng Sun, Brian Dillon, **Mohit Iyyer**. [How Much Do Modifications to Transformer Language Models Affect Their Ability to Learn Linguistic Knowledge?](#) *ACL Workshop on Insights from Negative Results in NLP*, 2022.

Dhruvil Gala, Mohammad Omar Khursheed, Hannah Lerner, Brendan O'Connor, **Mohit Iyyer**. [Analyzing Gender Bias within Narrative Tropes](#). *EMNLP Workshop on NLP and CSS*, 2020.

Anupam Guha, **Mohit Iyyer**, Jordan Boyd-Graber. [A Distorted Skull Lies in the Bottom Center... Identifying Paintings from Text Descriptions](#). *NAACL Human-Computer QA Workshop*, 2016.

Mohit Iyyer, Jordan Boyd-Graber, Hal Daumé III. [Generating Sentences from Semantic Vector Space Representations](#). *NIPS Workshop on Learning Semantics*, 2014.

PREPRINTS

Chaitanya Malaviya, Joseph Chee Chang, Dan Roth, **Mohit Iyyer**, Mark Yatskar, and Kyle Lo. [Contextualized Evaluations: Taking the Guesswork Out of Language Model Evaluations](#). *arXiv 2024*.

Shane Arora*, Marzena Karpinska*, Hung-Ting Chen, Ipsita Bhattacharjee, **Mohit Iyyer**, Eunsol Choi. [CaLMQA: Exploring culturally specific long-form question answering across 23 languages](#). *arXiv 2024*.

Rachneet Sachdeva, Yixiao Song, **Mohit Iyyer**, Iryna Gurevych. [Fine-grained Hallucination Detection and Mitigation in Long-form Question Answering](#). *arXiv 2024*.

Garima Dhanania, Sheshera Mysore, Chau Minh Pham, **Mohit Iyyer**, Hamed Zamani, Andrew McCallum. [Interactive Topic Models with Optimal Transport](#). *arXiv 2024*.

Simeng Sun, Dhawal Gupta, **Mohit Iyyer**. [Exploring the impact of low-rank adaptation on the performance, efficiency, and regularization of RLHF](#). *arXiv 2023*.

Pedro Rodriguez, Shi Feng, **Mohit Iyyer**, He He, Jordan Boyd-Graber. [Quizbowl: The Case for Incremental Question Answering](#). *arXiv 2019*.

FUNDING

In total, I have raised almost \$5M in research funding over the last five years from the NSF, industry labs, and philanthropic organizations.

- NSF RI MEDIUM, 2023-2027
 “Multilingual Long-form QA with Retrieval-Augmented Language Models”
 PI: Eunsol Choi, co-PI: Mohit Iyyer
UMass share: \$554,000

- Google Research Award, 2023
 “RankGen: Improving Text and Code Generation with Large Ranking Models”
 PI: Mohit Iyyer
\$30,000
- Kensho Research Award, 2023
 “Improving factuality via LLM alignment”
 PI: Mohit Iyyer
\$74,000
- NSF RETTL, 2022-2025
 “STEM Learning Embedded in a Machine-in-the-Loop Collaborative Story Writing Game”
 PI: Mohit Iyyer, co-PIs: Andrew Lan, Stephen Hood, Danielle McNamara
\$621,352
- Open Philanthropy, 2022-2024
 “Improving AI translation of novels into English”
 PI: Mohit Iyyer
\$822,365
- Alexa Prize TaskBot Challenge, 2021-2022
 PI: Hamed Zamani, co-PIs: Mohit Iyyer, Bruce Croft
\$250,000
- NSF CAREER, 2021-2026
 “Building Creative Writing Assistants for Machine-in-the-Loop Storytelling”
 PI: Mohit Iyyer
\$532,000
- NSF RI Medium, 2020-2024
 “Tree-Structured Self-Supervised Modeling for Natural Language”
 PI: Mohit Iyyer, co-PI: Andrew McCallum
\$1.2 million
- Subcontract with US Navy, 2020-2023
 “Natural language processing of fleet data”
 PI: Mohit Iyyer
\$136,295
- Adobe Faculty Research Gift, 2019-2023
 PI: Mohit Iyyer
\$200,000 cumulative
- IBM Faculty Award, 2020
 PI: Mohit Iyyer
\$20,000
- NSF CCRI Planning Grant, 2019-2020
 “Planning for the Development of a Platform to Support Multilingual and Multi-Domain Coreference Annotation for Natural Language Processing Research”
 PI: Brendan O’Connor, co-PI: Mohit Iyyer
\$100,000

- Genpact Faculty Gift, 2019
PI: Mohit Iyyer
\$50,000
- Sony Research Award, 2019
PI: Mohit Iyyer
\$60,000
- Intuit AI Award, 2019
PI: Mohit Iyyer
\$133,750
- Facebook Robust Deep Learning for NLP Award, 2019
PIs: Brendan O'Connor & Mohit Iyyer
\$76,905
- Allen Institute Young Investigator Award, 2018
PI: Mohit Iyyer
\$125,000

ADVISING

POSTDOCS:

Marzena Karpinska (2020-2024)

PHD ADVISEES:

Shufan Wang (2019-), CS

Katherine Thai (2020-), CS

Yixiao Song (2022-), Linguistics & CS, *co-advised with Rajesh Bhatt*

Yapei Chang (2022-), CS

Chau Pham (2022-), CS

Yekyung Kim (2023-), CS

Jenna Russell (2024-), CS

Rishanth Rajendhran (2024-), CS

PHD ALUMNI:

- Kalpesh Krishna (2018-2023), CS, *Google PhD Fellow, now research scientist at Google Gemini*
- Tu Vu (2018-2023), CS, *research scientist at Google DeepMind & Asst. Prof at Virginia Tech*
- Simeng Sun (2019-2024), CS, *research scientist at Nvidia*
- Andrew Drozdov (2018-2024, *co-advised with Andrew McCallum*), CS, *research scientist at Databricks*
- Nader Akoury (2018-2024), CS, *postdoc at Cornell*

ON COMMITTEE:

Emma Strubell (UMass CS PhD, 2019)
Patrick Verga (UMass CS PhD, 2019)
Carolyn Anderson (UMass Linguistics PhD, 2020)
Luke Vilnis (UMass CS PhD, 2020)
Su Lin Blodgett (UMass CS PhD, 2020)
Brandon Prickett (UMass Linguistics PhD, 2020)
Abe Handler (UMass CS PhD, 2021)
Katie Keith (UMass CS PhD, 2021)
Chen Qu (UMass CS PhD, 2021)
Chenyun Wu (UMass CS PhD, 2021)
Trapit Bansal (UMass CS PhD, 2021)
Rajarshi Das (UMass CS PhD, 2022)
Xiang Lorraine Li (UMass CS PhD, 2022)
Haw-Shiuan Chang (UMass CS PhD, 2022)
Subendhu Rongali (UMass CS PhD, 2022)
Difan Liu (UMass CS PhD, 2022)
Virat Shejwalkar (UMass CS PhD, 2022)
Jyoti Iyer (UMass Linguistics PhD, 2022)
Pinelopi Papalampidi (University of Edinburgh Informatics PhD, 2022)
Mahmood Jasim (UMass CS PhD, 2023)
Shahrazad Naseri (UMass CS PhD, 2023)
Pengshan Cai (UMass CS PhD, 2023)
Dung Thai (UMass CS PhD, 2023)
Pengshan Cai (UMass CS PhD, 2023)
Youngwoo Kim (UMass CS PhD, 2024)
Sheshera Mysore (UMass CS PhD, 2024)
Zhiqi Huang (UMass CS PhD, 2024)
Taisiya Glushkova (Instituto Superior Técnico CS PhD, 2024)

MENTORING:

Jiarui Liu (2023-), UMass CS MS
 Ipsita Bhattacharjee (2023-2024), UMass CS UG
 Ronan Salz (2022-2023), UMass CS UG
 Naveen Nizar (2022), UMass CS MS
 Nishant Raj (2022), UMass CS MS
 Xiaoyu Song (2021-2022), UMass CS MS
 George Wei (2021-2022), UMass CS UG
 Yapei Chang (2020-), Smith CS UG, *now PhD student in my group*
 Shriya Atmakuri (2021), UMass CS MS
 Evan Moore (2020-2021), UMass CS MS
 Kavya Jeganathan (2020-2021), UMass CS UG
 Luke Yeh (2018-2020), UMass CS UG
 Wenlong Zhao (2020-2021), UMass CS MS, *now CS PhD student at UMass*
 Sangeetha Balasubramanian (2020-2021), UMass CS MS, *now at ML engineer at Amazon*
 Akshita Bhagia (2019-2020), UMass CS MS *now research engineer at AI2*
 Dhruvil Gala (2019-2020), UMass CS UG, *now at Microsoft*
 Varun Sharma (2019-2020), UMass CS MS, *now at Bloomberg*
 Weiqiu You (2018-2020), UMass CS MS, *now PhD student at UPenn*
 Jack Merullo (2018-2020), UMass CS UG *now PhD student at Brown*
 Davis Yoshida (2016), CU Boulder undergraduate, *now PhD student at TTIC*
 Danny Bouman (2015), UMD College Park undergraduate

TEACHING

Spring 2024: Advanced Natural Language Processing (CS685), *204 students*
 Spring 2023: Advanced Natural Language Processing (CS685), *168 students*
 Spring 2022: Advanced Natural Language Processing (CS685), *110 students*
 Fall 2021: Advanced Natural Language Processing (CS685), *120 students*
 Fall 2020: Advanced Natural Language Processing (CS685), *122 students*
 Spring 2019: Deep Learning for Natural Language Processing (CS690D), *106 students, new course*
 Fall 2019: Introduction to Natural Language Processing (CS585), *200 students*
 Fall 2018: Introduction to Natural Language Processing (CS585), *107 students*

INVITED TALKS

Aug 2024: keynote at KDD CUP 2024 RAG workshop

Jul 2024: keynote at Workshop on Long Context Foundation Models (ICML 2024)

Jun 2024: talk at Tel Aviv University NLP seminar

Feb 2024: talk at University of Chicago / TTIC

Dec 2023: talk at University of Tokyo

Nov 2023: talk at MIT Embodied Intelligence seminar

Nov 2023: talk at UPenn CLunch

Oct 2023: talk at UMD CLIP Colloquium

Jul 2023: keynote at SIGIR Workshop on Retrieval-Enhanced Machine Learning

May 2023: talk at EACL Workshop on Insights from Negative Results in NLP

Apr 2023: talk at NYU Text-as-Data series

Dec 2022: talk at IndoML 2022

Jul 2022: lecture at 6th International Gran Canaria School on Deep Learning

May 2022: talk at Baidu Research (*remote*)

Apr 2022: talk at UNC Chapel Hill (*remote*)

Oct 2021: talk at Cornell AI seminar (*remote*)

Jun 2021: talk at Yandex NLP seminar (*remote*)

Jun 2021: talk at Cambridge NLP seminar (*remote*)

Mar 2021: talk at Georgia Tech NLP seminar (*remote*)

Sep 2020: talk at UMass CICS Computing and Social Justice series (*remote*)

Sep 2020: talk at UPenn CLunch (*remote*)

Aug 2020: talk at Data Science fwdays'20, Kyiv (*remote*)

Jul 2020: talk at USC/ISI Boston office (*remote*)

Nov 2019-Jan 2020: *Towards story generation*: NYU, Google AI Language, CMU LTI

Nov 2019: *Racial bias in sports commentary*: Brown University

Sep 2019: *Neural Question Answering and Generation*: IBM QA and Semantic Parsing workshop

Mar 2019: *Neural Language Modeling*: UMass Cognitive Brown Bag

Nov 2018: *Deep Learning for Question Answering*: University of Antwerp

Oct 2018-May 2019: *Towards Understanding Narratives with Artificial Intelligence*: University of Vermont (Science of Stories Symposium), WPI, UMass Data Science for the Humanities Series, Amazon Cambridge

July 2018: *Generating QA Dialogs from Documents*: TTIC Language Generation workshop

Mar 2018: *Generating Adversarial Examples with Syntactically Controlled Paraphrase Networks*: USC/ISI NL Seminar

Feb 2018: *Towards Understanding Creative Language with Artificial Intelligence*: Ursinus College

Jan 2018: *Generating Syntactic Adversaries with Controlled Paraphrasing*: Indian Institute of Science, Bengaluru

Feb-April 2017, *Deep Learning for Creative Language Understanding*: UC Davis, TTIC, CMU, Dartmouth, Northeastern, UC Santa Cruz, UMass Amherst, WUSTL, Microsoft Research, Allen Institute for AI

Jan 2017: Talk at CU Boulder Stats, Optimization, and Machine Learning seminar

Nov 2016: Talk at UMass Amherst Machine Learning & Friends Lunch

EXTERNAL SERVICE & OUTREACH

Senior Area Chair for ACL 2025, ACL 2024, ACL 2023, ACL 2022, COLING 2022, EMNLP 2021

Ethics Chair for COLM 2024

Area Chair for EMNLP 2024, ICLR 2024, EMNLP 2023, EMNLP 2022, ACL 2021, NAACL 2021, EMNLP 2020, ACL 2020, CoNLL 2019, EMNLP 2019

Program Committee/Reviewer for TACL, ACL, EMNLP, NAACL, ICLR, NIPS, ICML, CVPR

Sponsorship co-chair of NAACL 2024

Co-organizer of 5th Workshop on Narrative Understanding (WNU) at ACL 2023

Co-organizer of 4th Workshop on Narrative Understanding (WNU) at NAACL 2022

Co-organizer of 1st Workshop on Document-grounded Dialogue and Conversational QA (DialDoc) at ACL 2021

Co-organizer of 3rd Workshop on Narrative Understanding (WNU) at NAACL 2021

Co-organizer of 1st Joint Workshop on Narrative Understanding, Storylines, and Events (NUSE) at ACL 2020

Guest Editor of PLOS ONE [Science of Stories](#) Collection

Co-organizer of Narrative Understanding Workshop at NAACL 2019

Publicity Chair of EMNLP 2018

Co-organizer of Human-Computer Question Answering Competition at NIPS 2017

Co-organizer of Human-Computer Question Answering Workshop at NAACL 2016

Co-organizer of Mid-Atlantic Student Colloquium on Speech, Language and Learning (MASC-SLL) 2016

Talk at DC-NLP Meetup on “[Deep Learning for Question Answering](#)”

[Introductory talk](#) to high schoolers about the technology behind our quiz bowl system

INSTITUTIONAL SERVICE

Internal awards committee (2021-)

Graduate program committee (2020)

Data science faculty recruiting committee (2019)

Annual faculty review committee (2019)

Graduate program committee (2018)

PhD admissions committee (2018)

MEDIA

2024 TechCrunch [article](#) on our NoCha benchmark

2023 UMass Magazine [feature](#) on our literary machine translation work

2023 UMass [press release](#) covering our award-winning work on LongEval

Spoke about large language models and ChatGPT on local news stations in 2023: [GBH](#), [WWLP](#) (and again [here](#)), [Western Mass News](#) (and again [here](#))

EMNLP 2019 football racial bias press: [The Undefeated](#)

CVPR 2017 comic books press: [MIT Tech Review](#), [Digital Trends](#)

NAACL 2016 relationships press: [Aeon](#)

EMNLP 2014 quiz bowl press: [UMIACS](#), [Terp](#), [Diamondback](#), [Colorado CS](#)

QANTA [defeats Ken Jennings](#), 300-160

QANTA quiz bowl system [exhibition match](#) against a team of four former Jeopardy champions in May 2015 (result: 200-200 tie)

PATENTS

Richard Socher, Ankit Kumar, Ozan Irsoy, Mohit Iyyer, Caiming Xiong, Stephen Merity, Romain Paulus. *Dynamic memory network*. United States Patent #11113598. September 7, 2021.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer. *System and methods for performing NLP related tasks using contextualized word representations*. United States Patent #11030414. June 8, 2021.

EXHIBIT B

Materials Relied Upon

Academic Articles

1. [REDACTED].
2. Carolyn Caffrey Gardner & Gabriel J. Gardner, *Fast and Furious (at Publishers): The Motivations behind Crowdsourced Research Sharing*, 78 College & Research Libraries 131, 132 (2017), available at https://www.researchgate.net/publication/317117356_Fast_and_Furious_at_Publishers_The_Motivations_behind_Crowdsourced_Research_Sharing.
3. Vladimir V. Bochkarev et al., *The Average Word Length Dynamics as an Indicator of Cultural Changes in Society*, 12 Soc. Evolution & Hist. 153, 155 (2015), available at https://www.sociostudies.org/journal/files/seh/2015_2/153-175.pdf.

Produced Documents, Data and Source Code

1. ANT_BARTZ_000002544.
2. ANT_BARTZ_000436674.
3. ANT_BARTZ_000436675.
4. ANT_BARTZ_000004894.
5. ANT_BARTZ_000436666.
6. ANT_BARTZ_000000005.
7. ANT_BARTZ_000436661.
8. ANT_BARTZ_000436662.
9. ANT_BARTZ_000436663.
10. ANT_BARTZ_000436664.
11. ANT_BARTZ_000436665.
12. ANT_BARTZ_000436666.
13. BARTZ_ZHAO_CERT_000000506.
14. BARTZ_ZHAO_CERT_000000507.
15. BARTZ_ZHAO_CERT_000000509.

Deposition Transcripts

1. Deposition of Ben Y. Zhao.

Declarations

1. Declaration of Ben Y. Zhao, PhD in support of Plaintiffs' Motion for Class Certification.
2. Declaration of Mycal Tucker in support of Defendant Anthropic PBC's Opposition to Plaintiffs' Motion for Class Certification.

Web Sources

1. [REDACTED]
2. [REDACTED]
3. [REDACTED]
4. [REDACTED]
5. [REDACTED]
6. [REDACTED]
7. *Best Books of the Decade: 1920s*, goodreads.com, (last visited Apr. 15, 2025), available at https://www.goodreads.com/list/show/39.Best_Books_of_the_Decade_1920s.
8. *Best Books of the Decade: 1940s*, goodreads.com, (last visited Apr. 15, 2025), available at https://www.goodreads.com/list/show/23.Best_Books_of_the_Decade_1940s.
9. *Best Books of the Decade: 1970s*, goodreads.com, (last visited Apr. 15, 2025), available at <https://www.goodreads.com/list/show/21>.
10. [REDACTED]
11. Don Leeper, *How many pages will your book have? A castoff can tell you*, Bookmobile (July 13, 2015), available at <https://www.bookmobile.com/print-positive/how-many-pages-will-your-book-have/>.
12. [REDACTED]
13. [REDACTED]
14. *Non American Books that Every American Should Read*, goodreads.com (last visited Apr. 15, 2025), available at https://www.goodreads.com/list/show/5534.Non_American_books_that_every_American_should_read.
15. [REDACTED]



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28

CERTIFICATE OF SERVICE

I, Douglas A. Winthrop, am the ECF user whose identification and password are being used to file the foregoing **DECLARATION OF MOHIT IYER IN SUPPORT OF DEFENDANT ANTHROPIC PBC’S OPPOSITION TO PLAINTIFFS’ MOTION FOR CLASS CERTIFICATION.**

Dated: April 17, 2025

/s/ Douglas A. Winthrop