Annotated Bibliography on AI for Book-Lovers

Fred Zimmerman

Nimble Books LLC

https://nimblebooks.com/Annotated_Bibliography

Snapshot as of 9/3/2023; see website for latest

- # 4 Books So Powerful, They Can Rewire Your Brain

Item Type	Journal Article
Author	Anangsha Alammyan
Abstract	Positive psychology and neuroplasticity for a progressive mindset.
Date	2022-11-28T17:24:29.207Z
Language	en
URL	https://baos.pub/4-books-so-powerful-they-can-rewire-your-brain-40aa39e3d91c
Accessed	12/5/2022, 12:35:11 AM
Extra	WFZ I would argue that nearly *all* books are so powerful they can rewire your mind. This collection of four self-talk books is not a bad start. But the real opportunity is in our future deeper understanding of how particular books reshape the mind. Every book takes the reader through a somewhat predictable sequence of concepts, images, and emotions -- dozens or hundreds per minute -- with the detail of the neuropsychological reaction and its net effect mediated by the reader's personal experience. Eventually, books will be reverse engineered, going from the desired neurological outcomes backwards to the words.
Publication	Medium
Date Added	12/5/2022, 12:35:11 AM
Modified	12/5/2022, 2:39:14 AM

- **Attachments**
  - Snapshot

5 Recent, High Signal LLM Research Papers/Events

Item Type	Blog Post
Author	Bakz T. Future
Abstract	Everything in the Large Language Model/Multimodal AI world moves so fast. I try my best to keep up and surface these for others on my Twitter account. However, the problem with the high paced, flashy, twitter world, is that a lot of the most interesting papers get lost in the noise. This gives us no chance to revisit and reflect on the actual findings and implications.
Date	2022-12-19
URL	https://bakztfuture.substack.com/p/5-recent-high-signal-llm-research
Accessed	12/20/2022, 2:51:15 PM
Extra	WFZ Good roundup of several interesting papers. I'm skeptical of the Anthropic ones -- but the others are correctly advertised as potentially profound. H/T Bakz -- a great follow.
Blog Title	Multimodal by Bakz T. Future
Website Type	Substack newsletter
Date Added	12/20/2022, 2:51:15 PM

- **Attachments**

    o Snapshot

- # 5 Ways ChatGPT Can Improve, Not Replace, Your Writing

Item Type	Magazine Article
Author	David Nield
Abstract	Generate your own text—but get help from the AI bot to make it stand out.
Language	en-US
Library Catalog	www.wired.com
URL	https://www.wired.com/story/chatgpt-writing-tips/
Accessed	7/23/2023, 6:40:07 PM
Extra	Section: tags
Publication	Wired
ISSN	1059-1028
Date Added	7/23/2023, 6:40:07 PM

- **Tags:**

  - how-to
  - artificial intelligence
  - chatgpt

Attachments

  - Snapshot

- # 20B-parameter Alexa model sets new marks in few-shot learning

Item Type	Web Page
Abstract	With an encoder-decoder architecture — rather than decoder only — the Alexa Teacher Model excels other large language models on few-shot tasks such as summarization and machine translation.
Date	2022-08-02T12:59:36.259
Language	en
URL	https://www.amazon.science/blog/20b-parameter-alexa-model-sets-new-marks-in-few-shot-learning
Accessed	9/12/2022, 1:34:44 AM

Website Title	Amazon Science
Date Added	9/12/2022, 1:34:44 AM
Modified	9/12/2022, 1:34:47 AM

- **Attachments**

  - Snapshot

024_Euralex_2008_Maarten Janssen_Meaningless Dictionaries.pdf

Item Type	Journal Article
Author	Maarten Janssen
Date	2008
URL	https://euralex.org/elx_proceedings/Euralex2008/024_Euralex_2008_Maarten%20Janssen
Accessed	12/3/2022, 11:30:00 AM
Extra	WFZ He had me at 'meaningless dictionaries'. I love eccentric dictionaries. This is actually dictionaries that only try to do one thing per entry. This seems to be motivated by NLP ne context -- every function should have one job -- but also makes sense from the point of vie overload is a problem with both print and electronic dictionaries. I can definitely see short

Date Added	12/3/2022, 5:56:25 PM
Modified	12/5/2022, 12:21:11 AM

- # Attachments

  - 024_Euralex_2008_Maarten Janssen_Meaningless Dictionaries.pdf

---

- # '90s Dad Thrillers: a List

Item Type	Blog Post
Author	Max Read
Abstract	Notes toward a theory of the Dad Thriller
Date	2021-11-09
Short Title	'90s Dad Thrillers
URL	https://maxread.substack.com/p/90s-dad-thrillers-a-list
Accessed	6/17/2023, 7:37:23 PM
Blog Title	Read Max
Website Type	Substack newsletter

Date Added	6/17/2023, 7:37:23 PM
Modified	6/17/2023, 7:37:23 PM

- # [1909.08593] Fine-Tuning Language Models from Human Preferences

Item Type	Web Page
URL	https://arxiv.org/abs/1909.08593
Accessed	10/19/2022, 9:01:02 PM
Date Added	10/19/2022, 9:01:02 PM
Modified	10/19/2022, 9:01:07 PM

- ## Attachments

  - [1909.08593] Fine-Tuning Language Models from Human Preferences

- # [2207.02852] Machine Learning Model Sizes and the Parameter Gap

Item Type	Web Page
URL	https://arxiv.org/abs/2207.02852

Accessed	1/24/2023, 8:42:36 PM
Date Added	1/24/2023, 8:42:36 PM
Modified	1/24/2023, 8:42:40 PM

- ## Attachments

  - [2207.02852] Machine Learning Model Sizes and the Parameter Gap

- # [2303.17564] BloombergGPT: A Large Language Model for Finance

Item Type	Web Page
URL	https://arxiv.org/abs/2303.17564
Accessed	3/31/2023, 8:46:20 PM
Date Added	3/31/2023, 8:46:20 PM
Modified	3/31/2023, 8:46:24 PM

- ## Attachments

  - [2303.17564] BloombergGPT: A Large Language Model for Finance

- # [2304.08467] Learning to Compress Prompts with Gist Tokens

Item Type	Web Page
URL	https://arxiv.org/abs/2304.08467
Accessed	4/24/2023, 11:44:47 PM
Date Added	4/24/2023, 11:44:47 PM
Modified	4/24/2023, 11:44:47 PM

- # A Civil War Over Semicolons

Item Type	Web Page
Author	Gal Beckerman
Abstract	The biographer Robert Caro and his editor, Robert Gottlieb, have been arguing with each other for 50 years.
Date	2023-01-05T15:20:36Z
Language	en
URL	https://www.theatlantic.com/books/archive/2023/01/turn-every-page-documentary-robert-caro-robert-gottlieb/672651/
Accessed	1/11/2023, 7:41:49 PM

Extra	WFZ From the article: "But there's something else about the 50-yr relationship between LBJ biographer Robert Caro and copy editor Robert Gottlieb that gives a glimpse into another era: The two don't seem to like each other all that much. They bicker all the time, about every comma, period, and semicolon. Actually, don't even get them started on semicolons. Gottlieb refers to a "civil war" that took place over the punctuation mark's usage." My take: We will have these koind of curmudgeonly partnership relationships with our AIs. You can see this with ChatGPT, which is painfully pigheaded about forcing every sensitive topic into a three-paragraph explainer of "what is an LLM".
Website Title	The Atlantic
Date Added	1/11/2023, 7:41:49 PM
Modified	1/11/2023, 7:54:43 PM

- ## **Attachments**

  - ○ Snapshot

- # A Holistic Approach to Undesired Content Detection in the Real World

Item Type	Journal Article
Author	Todor Markov
Author	Chong Zhang
Author	Sandhini Agarwal

Author	Tyna Eloundou
Author	Teddy Lee
Author	Steven Adler
Author	Angela Jiang
Author	Lilian Weng

Abstract We present a holistic approach to building a robust and useful natural language classification system for real-world content moderation. The success of such a system relies on a chain of carefully designed and executed steps, including the design of content taxonomies and labeling instructions, data quality control, an active learning pipeline to capture rare events, and a variety of methods to make the model robust and to avoid overfitting. Our moderation system is trained to detect a broad set of categories of undesired content, including sexual content, hateful content, violence, self-harm, and harassment. This approach generalizes to a wide range of different content taxonomies and can be used to create high-quality content classifiers that outperform off-the-shelf models.

Date	2022-08-05
Library Catalog	DOI.org (Datacite)
URL	https://arxiv.org/abs/2208.03274
Accessed	8/9/2022, 11:57:43 PM
Rights	Creative Commons Attribution 4.0 International
Extra	WFZ OpenAI documents its content moderation process. tldr: lots of hands-on work with data and bespoke processes are required.
DOI	10.48550/ARXIV.2208.03274

Date Added	8/9/2022, 11:57:43 PM
Modified	12/4/2022, 11:55:40 PM

- ## Tags:

  - #nosource
  - Computation and Language (cs.CL)
  - FOS: Computer and information sciences
  - Machine Learning (cs.LG)

- # A hybrid model of complexity estimation: Evidence from Russian legal texts

Item Type	Conference Paper
Author	Olga Blinova
Author	Nikita Tarasov
Date	2022
Extra	WFZ Complexity estimation is an important topic for book publishing. This paper likely understates the complexity of the task because there is a gap between the written Russian law and its implementation.
Proceedings Title	Frontiers in artificial intelligence

- **Tags:**

  - #nosource

- # A New Frontier for Travel Scammers: A.I.-Generated Guidebooks

Item Type	Newspaper Article
Author	Seth Kugel
Author	Stephen Hiltner
Abstract	Shoddy guidebooks are flooding Amazon. Their authors claim to be renowned travel writers, but are they A.I. inventions? And how big is the problem?
Date	2023-08-05
Language	en-US
Short Title	A New Frontier for Travel Scammers
Library Catalog	NYTimes.com

URL	https://www.nytimes.com/2023/08/05/travel/amazon-guidebooks-artificial-intelligence.html
Accessed	8/7/2023, 9:26:05 PM
Extra	WFZ Numerous ethical problems here. 1) The low quality. 2) The intentional use of deceitful author biographies designed to trick buyers (real travel author Rick Steves, fictional travel author Mike Steves). 3) The low quality.
Section	Travel
Publication	The New York Times
ISSN	0362-4331
Date Added	8/7/2023, 9:26:05 PM
Modified	8/7/2023, 9:30:24 PM

- **Tags:**

  - Artificial Intelligence
  - Books and Literature
  - Computers and the Internet
  - Amazon.com Inc
  - Conde Nast Traveler
  - Consumer Reviews
  - E-Books and Readers
  - Federal Trade Commission
  - Fodor's
  - Frauds and Swindling
  - Frommer's

Attachments

- ○ Snapshot

A Novelist and an AI Cowrote Your Next Cringe-Read

Item Type	Magazine Article
Author	Kate Knibbs
Abstract	The new novel 'Amor Cringe' bills itself as "deepfake autofiction." WIRED talked to its coauthors, K. Allado-McDowell and the language model GPT-3.
Language	en-US
Library Catalog	www.wired.com
URL	https://www.wired.com/story/k-allado-mcdowell-gpt-3-amor-cringe/
Accessed	2/26/2023, 12:31:37 AM
Extra	WFZ Author K. Allado-McDowel takes an interesting tack in her novel AMOR CRINGE -- aiming to maximize "cringe". This feels like a strategy that is effective at working around some of the current limitations of the technology, but will not last. Eventually, and probably sooner rather than later, the strategy will shift simply to creating work of every kind that is of equal or higher quality than the very best human writing.

Publication	Wired
ISSN	1059-1028
Date Added	2/26/2023, 12:31:37 AM
Modified	2/28/2023, 2:49:04 AM

- **Tags:**

  - artificial intelligence
  - books

- ## A quantitative review of index compilation and indexing-related research in China in the twenty-first century (2000–17)

Item Type	Journal Article
Author	Wang Yanxiang
Date	2020
Library Catalog	Google Scholar
Volume	38
Pages	143–170

- ## Attachments

  - Snapshot

- # A survey of extract–transform–load technology

Item Type	Journal Article
Author	Panos Vassiliadis
Date	2009
Extra	WFZ Most highly cited article on ETL in Google Scholar.
Volume	5
Pages	1–27
Publication	International Journal of Data Warehousing and Mining (IJDWM)

Issue	3
Date Added	8/24/2022, 9:24:04 PM
Modified	12/4/2022, 11:55:43 PM

- **Tags:**

  - #nosource
  - ETL

- # A universal part-of-speech tagset

Item Type	Journal Article
Author	Dipanjan Das Slav Petrov
URL	http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.220.7779
Extra	Citation Key: SlavPetrov
Date Added	3/29/2021, 6:51:57 PM
Modified	12/4/2022, 11:52:44 PM

- **Tags:**

  - nlp

- pos
- part-of-speech
- tagset
- universal
- #nosource
- ⊖ No DOI found

A.I. Explore ::: Soviet Era propaganda posters

Item Type	Artwork
Artist	Brian Sykes
Abstract	Illustrations of famous movies and books using Midjourney prompts.
URL	https://www.linkedin.com/posts/brianwsykes_ai-creative-illustration-activity-6995010784871284736-_iIP?utm_source=share&utm_medium=member_desktop
Extra	WFZ My favorite image is the Soviet style poster of Mordor. https://tinyurl.com/yb2eh2hm
Date Added	11/6/2022, 8:41:39 PM
Modified	12/4/2022, 11:55:53 PM

Tags:

- #nosource

- # A.I. Is Mastering Language. Should We Trust What It Says?

Item Type	Newspaper Article
Author	Steven Johnson
Author	Nikita Iziev
Abstract	OpenAI's GPT-3 and other neural nets can now write original prose with mind-boggling fluency — a development that could have profound implications for the future.
Date	2022-04-15
Language	en-US
Library Catalog	NYTimes.com
URL	https://www.nytimes.com/2022/04/15/magazine/ai-language.html
Accessed	7/19/2022, 2:26:08 PM
Section	Magazine
Publication	The New York Times
ISSN	0362-4331
Date Added	7/19/2022, 2:26:08 PM

| **Modified** | 7/19/2022, 2:26:08 PM |

- **Tags:**

  - Artificial Intelligence
  - Software
  - Alphabet Inc
  - Computers and the Internet
  - Data-Mining and Database Marketing
  - DeepMind Technologies Ltd
  - Language and Languages
  - Meta Platforms Inc
  - OpenAI Labs
  - Writing and Writers

Attachments

  - Snapshot

- # Ablating Concepts in Text-to-Image Diffusion Models

Item Type	Preprint
Author	Nupur Kumari
Author	Bingliang Zhang

Author	Sheng-Yu Wang
Author	Eli Shechtman
Author	Richard Zhang
Author	Jun-Yan Zhu

Abstract   Large-scale text-to-image diffusion models can generate high-fidelity images with powerful compositional ability. However, these models are typically trained on an enormous amount of Internet data, often containing copyrighted material, licensed images, and personal photos. Furthermore, they have been found to replicate the style of various living artists or memorize exact training samples. How can we remove such copyrighted concepts or images without retraining the model from scratch? To achieve this goal, we propose an efficient method of ablating concepts in the pretrained model, i.e., preventing the generation of a target concept. Our algorithm learns to match the image distribution for a target style, instance, or text prompt we wish to ablate to the distribution corresponding to an anchor concept. This prevents the model from generating target concepts given its text condition. Extensive experiments show that our method can successfully prevent the generation of the ablated concept while preserving closely related concepts in the model.

Date	2023-03-23
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2303.13516
Accessed	4/7/2023, 4:02:33 AM
Extra	WFZ Useful for enforcing copyright, but I hope this technique does not spread widely. Dumbing down AI is the wrong way to go.
DOI	10.48550/arXiv.2303.13516

Repository	arXiv
Archive ID	arXiv:2303.13516
Date Added	4/7/2023, 4:02:33 AM
Modified	4/7/2023, 4:04:49 AM

- **Tags:**

  - Computer Science - Machine Learning
  - Computer Science - Computer Vision and Pattern Recognition
  - Computer Science - Graphics

Notes:

  - Comment: project website: https://www.cs.cmu.edu/~concept-ablation/

Attachments

  - arXiv Fulltext PDF
  - arXiv.org Snapshot

- # Absxcess #1

Item Type	Book

Author	Christopher English
Date	2022-21-10
Language	English
Library Catalog	Amazon
Extra	WFZ Manga produced using MidJourney art and Amazon KDP manga printing. https://amzn.to/3zsQugi
Date Added	10/28/2022, 4:04:45 PM
Modified	10/28/2022, 4:07:41 PM

- **Attachments**

  - Amazon.com Link

- # Actors say Hollywood studios want their AI replicas — for free, forever

Item Type	Web Page
Author	Andrew Webster
Abstract	No, this is not a Black Mirror episode.
Date	2023-07-13T20:35:09.191Z

Language	en-US
URL	https://www.theverge.com/2023/7/13/23794224/sag-aftra-actors-strike-ai-image-rights
Accessed	7/14/2023, 2:41:17 AM
Extra	WFZ The studios are taking a needlessly provocative stand: $200 for a one-day scan and unlimited rights forever. They are making IP companies look bad.
Website Title	The Verge
Date Added	7/14/2023, 2:41:17 AM
Modified	7/16/2023, 6:34:58 AM

- # Attachments

  - ○ Snapshot

- # Addressing "Documentation Debt" in Machine Learning Research: A Retrospective Datasheet for BookCorpus

Item Type	Preprint
Author	Jack Bandy

Author	Nicholas Vincent
Abstract	Recent literature has underscored the importance of dataset documentation work for machine learning, and part of this work involves addressing "documentation debt" for datasets that have been used widely but documented sparsely. This paper aims to help address documentation debt for BookCorpus, a popular text dataset for training large language models. Notably, researchers have used BookCorpus to train OpenAI's GPT-N models and Google's BERT models, even though little to no documentation exists about the dataset's motivation, composition, collection process, etc. We offer a preliminary datasheet that provides key context and information about BookCorpus, highlighting several notable deficiencies. In particular, we find evidence that (1) BookCorpus likely violates copyright restrictions for many books, (2) BookCorpus contains thousands of duplicated books, and (3) BookCorpus exhibits significant skews in genre representation. We also find hints of other potential deficiencies that call for future research, including problematic content, potential skews in religious representation, and lopsided author contributions. While more work remains, this initial effort to provide a datasheet for BookCorpus adds to growing literature that urges more careful and systematic documentation for machine learning datasets.
Date	2021-05-11
Short Title	Addressing "Documentation Debt" in Machine Learning Research
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2105.05241
Accessed	3/5/2023, 8:05:25 PM
Extra	arXiv:2105.05241 [cs]
DOI	10.48550/arXiv.2105.05241
Repository	arXiv

Archive ID	arXiv:2105.05241
Date Added	3/5/2023, 8:05:25 PM
Modified	3/5/2023, 8:05:25 PM

- **Tags:**

  - Computer Science - Computation and Language
  - Computer Science - Machine Learning
  - Computer Science - Computers and Society

Notes:

  - Comment: Working paper

Attachments

  - arXiv Fulltext PDF
  - arXiv.org Snapshot

- # Adversarial Attacks on Image Generation With Made-Up Words

Item Type	Journal Article
Author	Raphaël Millière

Abstract   Text-guided image generation models can be prompted to generate images using nonce words adversarially designed to robustly evoke specific visual concepts. Two approaches for such generation are introduced: macaronic prompting, which involves designing cryptic hybrid words by concatenating subword units from different languages; and evocative prompting, which involves designing nonce words whose broad morphological features are similar enough to that of existing words to trigger robust visual associations. The two methods can also be combined to generate images associated with more specific visual concepts. The implications of these techniques for the circumvention of existing approaches to content moderation, and particularly the generation of offensive or harmful images, are discussed.

Date   2022-08-04

Library Catalog   DOI.org (Datacite)

URL   https://arxiv.org/abs/2208.04135

Accessed   8/9/2022, 3:31:02 PM

Extra   WFZ Macaronic prompting, which involves designing cryptic hybrid words by concatenating subword units from different languages; and evocative prompting--leveraging morphological properties of existing words can be used to prime text to image models. While the article discusses this as a negative development in terms of evading content moderation, I am more positive. Secrets, shortcuts & cryptic languages have long played an important role in the formation of community, and books based on them tend to sell well! Cough, cough, "DaVinci Code."

DOI   10.48550/ARXIV.2208.04135

Date Added   8/9/2022, 3:31:02 PM

Modified   12/4/2022, 11:55:39 PM

- **Tags:**

  - #nosource
  - Computation and Language (cs.CL)
  - FOS: Computer and information sciences
  - Machine Learning (cs.LG)
  - Computer Vision and Pattern Recognition (cs.CV)
  - Cryptography and Security (cs.CR)

- # AG Introduces New Publishing Agreement Clauses Concerning AI

Item Type	Blog Post
Abstract	The Authors Guild is introducing four new model clauses concerning AI to its Model Trade Book Contract and Model Literary Translation Contract. In addition to the recent clause preventing the use of books in training generative AI without an author's […]
Language	en-US
URL	https://authorsguild.org/news/ag-introduces-new-publishing-agreement-clauses-concerning-ai/
Accessed	6/2/2023, 3:16:55 PM
Blog Title	The Authors Guild
Date Added	6/2/2023, 3:16:55 PM
Modified	6/2/2023, 3:16:55 PM

- **Attachments**

  - ○ Snapshot

- # AI and neurosensors combine to predict which songs will be hits

Item Type	Web Page
Author	Ryan Heath
Abstract	A study claims its technique achieves 97% accuracy.
Date	2023-06-27T09:45:00Z
Language	en
URL	https://www.axios.com/2023/06/27/ai-predicts-hits-number-one-songs
Accessed	6/28/2023, 6:12:49 AM
Website Title	Axios
Date Added	6/28/2023, 6:12:49 AM
Modified	6/28/2023, 6:12:49 AM

- **Attachments**

- # AI Faces Multiple Creator Lawsuits | CCC's Velocity of Content Podcast

Item Type	Blog Post
Author	https://velocityofcontentpodcast.com/author/rob
Abstract	Copyright infringement lawsuits from authors and artists against creators of Generative AI software are mounting.
Language	en-US
URL	https://velocityofcontentpodcast.com/ai-faces-multiple-creator-lawsuits/
Accessed	7/14/2023, 6:42:43 PM
Extra	WFZ More lawsuits. IMHO they will all be settled like Google's book scanning project.
Blog Title	Velocity of Content	A series of recordings from the Copyright Clearance Center
Date Added	7/14/2023, 6:42:43 PM
Modified	7/16/2023, 6:32:37 AM

- # AI is an opportunity for creative industries, says Bertelsmann boss

Item Type	Newspaper Article
Author	Olaf Storbeck
Author	Laura Pitel
Abstract	Thomas Rabe says that despite copyright challenges machine-generated content could be 'very positive'
Date	2023-05-25
URL	https://www.ft.com/content/1a02c92c-b5e6-454d-a303-44b7db04e306
Accessed	5/26/2023, 1:06:38 AM
Section	Artificial intelligence
Publication	Financial Times
Date Added	5/26/2023, 1:06:40 AM
Modified	5/26/2023, 1:06:40 AM

- # Attachments

  - ○ Snapshot

- # Ai Midjourney V4 Typography Explorations

Item Type	Web Page
Author	Alexandra Garcia Brillembourg
URL	https://www.linkedin.com/posts/alexandra-garcia-brillembourg-29812a5_midjourney-artificialinspiration-midjourney-activity-6994720857084354560-y_WV
Accessed	11/5/2022, 3:49:44 PM
Extra	WFZ Beautiful use of generative AI in typography.
Website Title	Ai Midjourney V4 Typography Explorations
Date Added	11/5/2022, 3:49:44 PM
Modified	11/5/2022, 3:56:17 PM

- # **Attachments**

  - ○ Post | Feed | LinkedIn

- # AI narration is inevitable

Item Type	Web Page
Abstract	Publishers need to embrace the exciting potential of AI narration for audiobooks.
Language	En

URL	https://www.thebookseller.com/comment/ai-narration-is-inevitable
Accessed	1/13/2023, 2:12:31 AM
Website Title	The Bookseller
Date Added	1/13/2023, 2:12:31 AM
Modified	1/13/2023, 2:12:34 AM

- **Attachments**

  - Snapshot

- # Ai Papercut Typography Artificial Intelligence

Item Type	Web Page
Author	Alexandra Garcia Brillembourg
Abstract	Graphic Design,Adobe Photoshop,Midjourney
Date	2022-1206
Language	english
URL	https://www.behance.net/gallery/157957055/Ai-Papercut-Typography-Artificial-Intelligence

Accessed	12/6/2022, 10:14:23 PM
Extra	WFZ Paper-cut books are a venerable niche and artform in themselves! See, e.g. the Guild of American Papercutters at https://papercutters.org/books-and-starter-kits. See also Kreisel 1994 for bibliography of the history of papercut books (which covers many centuries and cultures).
Website Title	Behance
Date Added	12/6/2022, 10:14:23 PM
Modified	12/6/2022, 10:25:59 PM

- ## **Attachments**
  - ○ Snapshot

- # ai_policy_guidance.pdf

Item Type	Web Page
URL	https://www.copyright.gov/ai/ai_policy_guidance.pdf
Accessed	6/19/2023, 7:09:30 PM
Extra	WFZ Copyright guidance on AI produced material. The key point is that the more human involvement in the final product, the more likely a work is to be copyrightable.
Date Added	6/19/2023, 8:58:11 PM

Modified	6/19/2023, 9:00:11 PM

- ## **Attachments**

  - ai_policy_guidance.pdf

- # Alan Dean Foster Fan Page | Facebook

Item Type	Web Page
Author	Alan Dean Foster
Date	2022
URL	https://www.facebook.com/AlanDeanFoster
Accessed	9/9/2022, 10:10:39 PM
Extra	WFZ [Alan Dean Foster is a science fiction author](https://amzn.to/3B8eC7Y) best known for his Flinx and Commonwealth series and his numerous high-profile movie adaptations (IN THE SPLINTER OF THE MIND'S EYE, ALIEN ...) This is what happens when an extremely prolific and talented author gets his hands on DALL-E. He illustrates his world! [Alan Dean Foster images of MidWorld and Prism](https://www.facebook.com/AlanDeanFoster/photos) (Facebook group, you may have to login/join) Now imagine every author doing this as a matter of course!
Date Added	9/9/2022, 10:10:39 PM
Modified	9/23/2022, 3:26:25 PM

- **Attachments**

  - Alan Dean Foster Fan Page | Facebook

- # AlphaDev discovers faster sorting algorithms

Item Type	Web Page
Abstract	In our paper published today in Nature, we introduce AlphaDev, an artificial intelligence (AI) system that uses reinforcement learning to discover enhanced computer science algorithms – surpassing those honed by scientists and engineers over decades.
Language	en
URL	https://www.deepmind.com/blog/alphadev-discovers-faster-sorting-algorithms?utm_source=twitter&utm_medium=social&utm_campaign=OCS
Accessed	6/8/2023, 3:32:55 AM
Date Added	6/8/2023, 3:32:55 AM
Modified	6/8/2023, 3:32:55 AM

- **Attachments**

  - Snapshot

- # aman.ai • exploring the art of artificial intelligence

Item Type	Web Page
URL	https://aman.ai/
Accessed	4/9/2023, 4:04:26 PM
Extra	WFZ Exemplary reviews of basic concepts in AI and other fields in a lucid, compact style that is itself something of an innovation in technical writing.
Date Added	4/9/2023, 4:04:26 PM
Modified	4/9/2023, 4:05:50 PM

- # **Attachments**

  - aman.ai • exploring the art of artificial intelligence

- # Amazon Search: The Joy of Ranking Products

Item Type	Conference Paper
Author	Daria Sorokina
Author	Erick Cantu-Paz

Abstract	Amazon is one of the world's largest e-commerce sites and Amazon Search powers the majority of Amazon's sales. As a consequence, even small improvements in relevance ranking both positively influence the shopping experience of millions of customers and significantly impact revenue. In the past, Amazon's product search engine consisted of several handtuned ranking functions using a handful of input features. A lot has changed since then. In this talk we are going to cover a number of relevance algorithms used in Amazon Search today. We will describe a general machine learning framework used for ranking within categories, blending separate rankings in All Product Search, NLP techniques used for matching queries and products, and algorithms targeted at unique tasks of specific categories — books and fashion.
Date	2016-07-07
Language	en
Short Title	Amazon Search
Library Catalog	DOI.org (Crossref)
URL	https://dl.acm.org/doi/10.1145/2911451.2926725
Accessed	9/1/2022, 10:54:38 AM
Extra	WFZ From a book-publishing perspective, most interesting nugget is that the goal of product search is to capture the full range of valid product type and product descriptions. So, genre and microgenre may be especially worth emphasizing in keyword generation. "We use detected product types in queries and product descriptions to create powerful features in ranking models."
Place	Pisa Italy
Publisher	ACM
ISBN	978-1-4503-4069-4

Pages	459-460
Proceedings Title	Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval
Conference Name	SIGIR '16: The 39th International ACM SIGIR conference on research and development in Information Retrieval
DOI	10.1145/2911451.2926725
Date Added	9/1/2022, 10:54:38 AM
Modified	9/1/2022, 11:01:41 AM

- # Attachments

  - Sorokina and Cantu-Paz - 2016 - Amazon Search The Joy of Ranking Products.pdf

- # Amazon. com recommendations: Item-to-item collaborative filtering

Item Type	Journal Article
Author	Greg Linden
Author	Brent Smith
Author	Jeremy York
Date	2003

- **Attachments**

  - Full Text
  - Snapshot

- AMPTP Disputes SAG-AFTRA's "Misleading" Claims About Last Contract Offer Before Strike Began; Union Responds – Update

Item Type	Blog Post
Author	David Robb
Abstract	UPDATED with SAG-AFTRA response: The Alliance of Motion Picture and Television Producers is disputing SAG-AFTRA's "misleading" characterization of its proposals before contract talks broke off and …
Date	2023-07-21T21:29:51+00:00
Language	en-US
URL	https://deadline.com/2023/07/actors-strike-studios-dispute-sag-aftra-contract-offer-claims-1235445314/
Accessed	7/23/2023, 1:10:23 AM
Blog Title	Deadline
Date Added	7/23/2023, 1:10:23 AM
Modified	7/23/2023, 1:10:23 AM

- ## **Notes:**

  - o We need a balanced approach based on careful use, not prohibition. Among other protections, the proposal provides that Producers: ● Must obtain a background actor's consent to use a "digital replica" other than for the motion picture for which the background actor was hired. Producers told SAG-AFTRA they would

agree to apply the same provisions that the Producers proposed would apply to performers, so that consent and separate bargaining for payment must take place at the time of use. ● Cannot use "digital replicas" of background actors in lieu of hiring the required number of covered background actors under the Agreement. ● Must obtain a performer's consent to create a "digital replica" for use in a motion picture. ● Must obtain a performer's consent to digitally alter the performance beyond typical alterations that have historically been done in post-production. ● Must obtain a performer's consent and bargain separately for use of a "digital replica" other than for the motion picture for which the performer was hired. ● Producers told SAG-AFTRA they would agree to SAG-AFTRA's proposal that consent to use a "digital replica" must include a "description of the intended use." Likewise, consent to digital alterations must include a "description of the intended alterations."

- # An Open Letter to AI

Item Type	Blog Post
Author	Fiverr.
URL	https://www.facebook.com/photo/?fbid=10160422907308563&set=a.122293203562
Accessed	1/29/2023, 2:24:47 PM

Extra	WFZ Fiverr. smartly embraces selling AI-enhanced services. Unfortunately, true to Fiverr's brand, most of the "gigs" offered are around $20/shot. Not an encouraging sign for those hoping to prosper from AI.
Blog Title	Lewis Shepherd's Facebook Page
Date Added	1/29/2023, 2:24:47 PM
Modified	1/29/2023, 11:30:54 PM

- ## Attachments
  - Facebook

- # An Overview of the Tesseract OCR Engine

Item Type	Conference Paper
Author	Ray Smith
Date	2007
Library Catalog	Google Research
Pages	629–633
Proceedings Title	Proc. Ninth Int. Conference on Document Analysis and Recognition (ICDAR)

Date Added	9/17/2022, 7:51:38 PM
Modified	9/17/2022, 7:51:46 PM

- **Tags:**
  - ⊖ No DOI found

Attachments

  - Full Text PDF

- # Analysis of a Play by Means of CHAPLIN, the Characters and Places Interaction Network Software

Item Type	Journal Article
Author	Amelia Carolina Sparavigna
Abstract	Recently, we have developed a software able of gathering information on social networks from written texts. This software, the CHAracters and PLaces Interaction Network (CHAPLIN) tool, is implemented in Visual Basic. By means of it, characters and places of a literary work can be extracted from a list of raw words. The software interface helps users to select their names out of this list. Setting some parameters, CHAPLIN creates a network where nodes represent characters/places and edges give their interactions. Nodes and edges are labelled by performances. In this paper, we propose to use CHAPLIN for the analysis a William Shakespeareâ€™s play, the famous â€œTragedy of Hamlet, Prince of Denmarkâ€ . Performances of characters in the play as a whole and in each act of it are given by graphs.

- **Tags:**

  - #nosource

- # Analysis of Structure and Plots of Characters from Plays and Novels to Create Novel and Plot Data Bank (NPDB)

- **Tags:**

  - #nosource

- # Anthropic's $5B, 4-year plan to take on OpenAI

Item Type	Blog Post
Author	Kyle Wiggers
Abstract	AI research startup Anthropic aims to raise as much as $5 billion over the next two years to take on rival OpenAI.
Date	2023-04-06T21:25:11+00:00
Language	en-US
URL	https://techcrunch.com/2023/04/06/anthropics-5b-4-year-plan-to-take-on-openai/
Accessed	4/22/2023, 5:01:29 PM
Extra	WFZ Anthopic believes its 2025-26 model will be able to generate complete books, screenplays, videos, and songs. \nArticle says: This frontier model could be used to build virtual assistants that can answer emails, perform research and generate art, books and more, some of which we have already gotten a taste of with the likes of GPT-4 and other large language models. "These models could begin to automate large portions of the economy," the pitch deck reads. "We believe that companies that train the best 2025/26 models will be too far ahead for anyone to catch up in subsequent cycles."
Blog Title	TechCrunch
Date Added	4/22/2023, 5:01:26 PM
Modified	4/22/2023, 5:05:37 PM

- # Notes:

- This frontier model could be used to build virtual assistants that can answer emails, perform research and generate art, books and more, some of which we have already gotten a taste of with the likes of GPT-4 and other large language models. "These models could begin to automate large portions of the economy," the pitch deck reads. "We believe that companies that train the best 2025/26 models will be too far ahead for anyone to catch up in subsequent cycles."

- # Anti-Plagiarism Software Market Size Will Achieve USD 4,806 Million By 2030 Growing at 23.3% CAGR - Exclusive Report By Acumen Research and Consulting

Item Type	Web Page
Author	Acumen Research and Consulting
Abstract	Acumen Research and Consulting recently published report titled "Anti-Plagiarism Software Market Size, Share, Analysis Report and Region Forecast, 2022 -...
Date	8/21/2022 10:00:00 PM
Language	en
URL	https://www.globenewswire.com/news-release/2022/08/21/2501830/0/en/Anti-Plagiarism-Software-Market-Size-Will-Achieve-USD-4-806-Million-By-2030-

Accessed	10/8/2022, 4:02:25 PM
Extra	WFZ Marketing research firm puts some numbers around the size and expected rapid growth of anti-plagiarism service industry: $4.8B by 2030. Sadly, this seems likely correct as most industries, including publishing, education, and media, have barely begun to perceive the coming impact of large language models.
Website Title	GlobeNewswire News Room
Date Added	10/8/2022, 4:02:25 PM
Modified	10/8/2022, 6:59:50 PM

- ## **Attachments**

  - Snapshot

- # Artificial ASMR: A Cyber-Psychological Study

Item Type	Preprint
Author	Zexin Fang
Author	Bin Han
Author	C. Clark Cao

Author	Hans D. Schotten
Abstract	The popularity of Autonomous Sensory Meridian Response (ASMR) has skyrockteted over the past decade, but scientific studies on it are still few and immature. With our attention caught by the common acoustic patterns in ASMR audios, we investigate the correlation between the time-frequency and cyclic features of audio signals and their effectiveness in triggering ASMR effects. A cyber-psychological approach that combines signal processing, artificial intelligence, and experimental psychology is taken, with which we are able to identify ASMR-related acoustic features, and therewith synthesize random artificial ASMR audios.
Date	2022-10-27
Language	en
Short Title	Artificial ASMR
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2210.14321
Accessed	12/8/2022, 2:12:00 AM
Extra	WFZ Machine language techniques for creating ASMR trigger audio. I see no reason in principle why these techniques could not be used to create ASMR for book-related triggers.
Repository	arXiv
Archive ID	arXiv:2210.14321
Date Added	12/8/2022, 2:12:00 AM
Modified	12/8/2022, 2:12:58 AM

- **Tags:**

  - Computer Science - Artificial Intelligence
  - Computer Science - Multimedia
  - Computer Science - Sound
  - Electrical Engineering and Systems Science - Audio and Speech Processing
  - Electrical Engineering and Systems Science - Signal Processing

Notes:

  - Comment: Submitted to ICASSP 2023

Attachments

  - Fang et al. - 2022 - Artificial ASMR A Cyber-Psychological Study.pdf

Artificial Intelligence and Copyright

Item Type	Web Page
Abstract	The United States Copyright Office is undertaking a study of the copyright law and policy issues raised by artificial intelligence (``AI'') systems. To inform the Office's study and help assess whether legislative or regulatory steps in this area are warranted, the Office seeks comment on these...
Date	2023-08-30

URL	https://www.federalregister.gov/documents/2023/08/30/2023-18624/artificial-intelligence-and-copyright
Accessed	9/2/2023, 1:29:19 PM
Extra	WFZ Crucial request for comment from Copyright Office that gives an excellent survey of the issues they are considering both from the side of generative AI creators and from the side of copyright holders.
Website Title	Federal Register
Date Added	9/2/2023, 1:29:19 PM
Modified	9/3/2023, 5:58:16 AM

- ## Attachments

  - ○ Snapshot

- # Ask My Book: The Minimalist Entrepreneur

Item Type	Web Page
Author	Sahil Lavignia
Abstract	Ask questions of my book, get answers in my voice, powered by AI
URL	https://askmybook.com

Accessed	11/15/2022, 5:07:18 PM
Extra	WFZ Ask questions and the book responds in the author's voice. Every book should and likely will have this feature in future. People always talk about books being in conversation with one another. It may happen at scale, many:many, faster than we think: a Babel of books. Will they understand one another?
Website Title	Ask My Book
Date Added	11/15/2022, 5:07:18 PM
Modified	11/16/2022, 2:49:17 AM

- ## Attachments

  - Snapshot

- # [asmr books] - YouTube

Item Type	Web Page
Author	Various YouTube Creators
Abstract	YouTube answer set for a search on [ASMR books].
URL	https://www.youtube.com/results?search_query=asmr+books
Accessed	12/8/2022, 12:40:47 AM
Extra	WFZ Autonomous Sensory Meridian Response (ASMR) is a video & audio technique for triggering a goosebumps response in viewers and listeners. Almost

	anything can be an ASMR trigger -- including stimuli related to the sensory experience of reading books! See also Fang 2022 supra.
Website Type	Web
Website Title	YouTube
Date Added	12/8/2022, 12:40:47 AM
Modified	12/8/2022, 5:27:18 AM

- ## Attachments

  - ## asmr books - YouTube

- # Associative Classification in Text Categorization

Item Type	Document
Author	Jian Chen
Author	Jian Yin
Author	Jun Zhang
Author	Jin Huang
URL	http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.140.4886

Extra	Citation Key: Chen
Date Added	3/29/2021, 6:52:42 PM
Modified	12/4/2022, 11:53:35 PM

- # Tags:

  - o #nosource

- # Asterix Writer

Item Type	Web Page
Abstract	Edit your writing, like you edit a photo
URL	https://www.asterixwriter.com/features
Accessed	3/12/2023, 6:58:06 PM
Website Title	Asterix
Date Added	3/12/2023, 6:58:06 PM
Modified	3/12/2023, 6:58:06 PM

- # Attachments

Attention Is All You Need

Item Type	Preprint
Author	Ashish Vaswani
Author	Noam Shazeer
Author	Niki Parmar
Author	Jakob Uszkoreit
Author	Llion Jones
Author	Aidan N. Gomez
Author	Lukasz Kaiser
Author	Illia Polosukhin

Abstract — The dominant sequence transduction models are based on complex recurrent or convolutional neural networks in an encoder-decoder configuration. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show

that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Date	2017-12-05
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/1706.03762
Accessed	8/24/2022, 3:27:35 AM
Extra	arXiv:1706.03762 [cs]
DOI	10.48550/arXiv.1706.03762
Repository	arXiv
Archive ID	arXiv:1706.03762
Date Added	8/24/2022, 3:27:35 AM
Modified	8/24/2022, 3:27:35 AM

- ## Tags:

  - Computer Science - Computation and Language
  - Computer Science - Machine Learning

Notes:

- Comment: 15 pages, 5 figures

Attachments

Authors are losing their patience with AI, part 349235

Item Type	Blog Post
Author	Amanda Silberling
Abstract	Prosecraft creator Benji Smith took down his database of 27,000 books, which were scraped without authors' consent.
Date	2023-08-07T22:09:45+00:00
Language	en-US
URL	https://techcrunch.com/2023/08/07/authors-ai-prosecraft/
Accessed	8/8/2023, 2:41:29 PM
Extra	WFZ Sad. Related item in this bibliography has author's explanation.
Blog Title	TechCrunch
Date Added	8/8/2023, 2:41:29 PM
Modified	8/9/2023, 6:53:45 PM

- **Attachments**
  - Snapshot

Authorship and AI tools

Item Type	Web Page
Abstract	The COPE position statment on the use of AI tools in research publications states that AI tools cannot be listed as an author of a paper as they cannot take responsibility for submitted work.
Language	en
URL	https://publicationethics.org/cope-position-statements/ai-author
Accessed	8/29/2023, 6:31:30 AM
Website Title	COPE: Committee on Publication Ethics
Date Added	8/29/2023, 6:31:30 AM
Modified	8/29/2023, 6:31:30 AM

- **Attachments**
  - Snapshot

- # Automatic Personalized Text Summarization Agent using Generic Relevance Weight based on NMF

Item Type	Journal Article
Author	Sun Park
Extra	Citation Key: Park
Pages	1-3
Date Added	3/29/2021, 6:50:59 PM
Modified	12/4/2022, 11:51:31 PM

- ## Tags:

  - #nosource

- # Automatic text summarization methods: A comprehensive review

Item Type	Journal Article
Author	Grishma Sharma
Author	Deepak Sharma

- **Tags:**

  - text summarization

- # Automatic Text Summarization Using Two-Step

Item Type	Journal Article
Author	Wooncheol Jung
Author	Youngjoong Ko
Author	Jungyun Seo
Extra	Citation Key: Jung

Pages	71-81
Date Added	3/29/2021, 6:52:02 PM
Modified	12/4/2022, 11:52:49 PM

- # Tags:

  - #nosource

- # AutoMuse

Item Type	Web Page
Author	David Shapiro
Abstract	What is AutoMuse? AutoMuse started as a collection of experiments with the latest, most powerful AI models available. As a writer, I wanted to see what I could get these AI engines to do, so I started monkeying around. Dozens of experiments later, I realized I had a veritable treasure trove of
Language	en-US
URL	https://www.automuse.io/home
Accessed	10/8/2022, 10:28:47 AM
Extra	WFZ David Shapiro has been a leading contributor to the OpenAI community since early days. He has written some very useful tools including a Recursive Summarizer that I forked. He has scaled up his ambitions in an interesting way!

Date Added	10/8/2022, 10:28:47 AM
Modified	12/4/2022, 11:55:47 PM

- **Tags:**
  - #nosource

Axios interview: Zients says Biden will push for more AI authority

Item Type	Web Page
URL	https://www.axios.com/2023/07/21/zients-biden-ai-authority?utm_source=newsletter&utm_medium=email&utm_campaign=newsletter_axio
Accessed	7/21/2023, 6:45:35 PM
Extra	WFZ "How rigorous are your security protocols?" Zients said. "We want to get watermark that an image or text passage is AI-generated] in place. How fast does that happen?" I don watermarking: a major invasion of user privacy.
Date Added	7/21/2023, 6:45:35 PM
Modified	7/21/2023, 9:35:24 PM

- **Attachments**

- - Axios interview: Zients says Biden will push for more AI authority

- # Bag of Tricks for Efficient Text Classification

Item Type	Preprint
Author	Armand Joulin
Author	Edouard Grave
Author	Piotr Bojanowski
Author	Tomas Mikolov
Abstract	This paper explores a simple and efficient baseline for text classification. Our experiments show that our fast text classifier fastText is often on par with deep learning classifiers in terms of accuracy, and many orders of magnitude faster for training and evaluation. We can train fastText on more than one billion words in less than ten minutes using a standard multicore~CPU, and classify half a million sentences among~312K classes in less than a minute.
Date	2016-08-09
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/1607.01759
Accessed	8/29/2022, 5:58:23 AM
Extra	WFZ I switched to this because it is ~ 100x faster than alternatives. https://stackoverflow.com/a/62115619/1024200

- ## Tags:

  - Computer Science - Computation and Language

Attachments

  - **arXiv Fulltext PDF**
  - **arXiv.org Snapshot**

- # Batch Prompting: Efficient Inference with Large Language Model APIs

Item Type	Preprint
Author	Zhoujun Cheng
Author	Jungo Kasai
Author	Tao Yu

Abstract	Performing inference on hundreds of thousands of samples with large language models (LLMs) can be computationally and financially costly. We propose batch prompting, a simple alternative prompting approach that enables the LLM to run inference in batches, instead of one sample at a time. Our method reduces both token and time costs while retaining downstream performance. We theoretically demonstrate that under a few-shot in-context learning setting, the inference costs decrease almost inverse linearly with the number of samples in each batch. We extensively validate the effectiveness of batch prompting on ten datasets across commonsense QA, arithmetic reasoning, and NLI/NLU: batch prompting significantly~(up to $5\times$ with six samples in batch) reduces the LLM (Codex) inference token and time costs while achieving better or comparable performance. Our analysis shows that the number of samples in each batch and the complexity of tasks affect its performance. Further, batch prompting can be applied across different LLMs and reasoning methods.
Date	2023-01-18
Short Title	Batch Prompting
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2301.08721
Accessed	1/24/2023, 1:28:43 PM
Extra	arXiv:2301.08721 [cs] version: 1
DOI	10.48550/arXiv.2301.08721
Repository	arXiv
Archive ID	arXiv:2301.08721
Date Added	1/24/2023, 1:28:43 PM
Modified	1/24/2023, 1:28:43 PM

- **Tags:**

  - Computer Science - Computation and Language
  - Computer Science - Artificial Intelligence

Notes:

  - Comment: 18 pages, 9 figures

Attachments

  - arXiv Fulltext PDF
  - arXiv.org Snapshot

- # Beauty Is in the Eye of the Beholder— but Memorability May Be Universal

Item Type	Magazine Article
Author	Celia Ford
Abstract	When humans and a neural network viewed pieces of art, they all found the same images memorable. What those images have in common offers a glimpse into what fascinates the brain.
Language	en-US
Library Catalog	www.wired.com

- ## Tags:

    - artificial intelligence
    - art
    - brains and behavior
    - deep learning
    - memory
    - neural networks
    - neuroscience
    - perception
    - psychology

Attachments

    - Snapshot

- # BibleGPT

Item Type	Web Page
URL	https://biblegpt.org/
Accessed	1/26/2023, 8:04:19 PM
Extra	WFZ Talk to the Book. I tried asking it a straightforward theological question of low difficulty -- "what are principalities?" -- and it did well. Didn't try more adversarial approaches--not sure if that wouldn't be blasphemy! Every book should be able to talk.
Date Added	1/26/2023, 8:04:19 PM
Modified	1/27/2023, 5:16:51 AM

- ## Attachments

  - BibleGPT

- # Biblioracle: Can an android write a book? A new app gets scary close.

Item Type	Web Page
Abstract	But what if you could outsource the writing of a book to an algorithm, and the book could be drafted within a matter of minutes?
Language	English (United States), en-US

Short Title	Biblioracle
URL	https://www.chicagotribune.com/entertainment/books/ct-ent-biblioracle-algorithm-write-books-20220528-tqo7jcpmsng2blm37igcpl7saa-story.html
Accessed	7/19/2022, 2:25:04 PM
Website Title	Chicago Tribune
Date Added	7/19/2022, 2:25:04 PM
Modified	7/19/2022, 2:25:04 PM

- ## **Attachments**

  - ○ Snapshot

- # Big Bird : Transformers for Longer Sequences

Item Type	Journal Article
Author	Manzil Zaheer
Author	Avinava Dubey
Author	Joshua Ainslie
Author	Chris Alberti

Author   Santiago Ontanon

Author   Philip Pham

Author   Anirudh Ravula

Author   Qifan Wang

Author   Li Yang

Author   Amr Ahmed

Author   A Nystr

Author   Zhanpeng Zeng

Author   Rudrasis Chakraborty

Author   Glenn Fung

Author   Yin Li

Author   Vikas Singh

Author   Mingxing Tan

Date Added	3/29/2021, 6:54:14 PM
Modified	12/4/2022, 11:55:01 PM

- ## Tags:

  - #nosource

- # BLEU: a method for automatic evaluation of machine translation

Item Type	Conference Paper
Author	Kishore Papineni
Author	Salim Roukos
Author	Todd Ward
Author	Wei-Jing Zhu
Abstract	Human evaluations of machine translation are extensive but expensive. Human evaluations can take months to finish and involve human labor that can not be reused. We propose a method of automatic machine translation evaluation that is quick, inexpensive, and language-independent, that correlates highly with human evaluation, and that has little marginal cost per run. We present this method as an automated understudy to skilled human judges which substitutes for them when there is need for quick or frequent evaluations.
Date	July 6, 2002

Short Title	BLEU
Library Catalog	ACM Digital Library
URL	https://doi.org/10.3115/1073083.1073135
Accessed	9/15/2022, 8:00:00 PM
Extra	WFZ The BLEU bilingual evaluation understudy is a measure from 0 to 1 of the quality of machine translation, with 1 being top-notch human professional translation. We need similar metrics for writing quality.
Place	USA
Publisher	Association for Computational Linguistics
Pages	311–318
Series	ACL '02
Proceedings Title	Proceedings of the 40th Annual Meeting on Association for Computational Linguistics
DOI	10.3115/1073083.1073135
Date Added	9/17/2022, 2:28:28 AM
Modified	9/17/2022, 3:10:26 AM

- # Attachments

  - Full Text PDF

- # Book Cover Dataset

Item Type	Software
Programmer	Iwana, Brian Kenji
Abstract	This dataset contains 207,572 books from the Amazon.com, Inc. marketplace.
Date	2022-07-24T11:38:57Z
Library Catalog	GitHub
URL	https://github.com/uchidalab/book-dataset
Accessed	8/6/2022, 5:57:32 AM
Rights	Apache-2.0
Extra	WFZ CSV file with basic info about 207,000 Amazon titles in 32 genres including thumbnail images.
Company	九州大学 ヒューマンインタフェース研究室
Prog. Language	Python
Date Added	8/6/2022, 5:57:32 AM
Modified	12/4/2022, 11:55:39 PM

- ## Tags:

-

- # Book Indexing Rates

Item Type	Blog Post
Author	Index Busters
Abstract	Numerous factors can influence book indexing rates. Indexing fees can be calculated per hour/page/word. Average rates are usually between USD $2.50-6.50
Language	en-US
URL	https://www.indexbusters.com/book-indexing-rates/
Accessed	8/22/2022, 3:47:05 PM
Extra	WFZ Representative company fee schedule.
Blog Title	Index Busters
Date Added	8/22/2022, 3:47:05 PM
Modified	8/22/2022, 3:48:32 PM

- ## Attachments

- o Snapshot

- # Book Parts

Item Type	Book
Author	Dennis Duncan
Author	Adam Smyth

Abstract

What would an anatomy of the book look like? There is the main text, of course, the file that the author proudly submits to their publisher. But around this, hemming it in on the page or enclosing it at the front and back of the book, there are dozens of other texts—page numbers and running heads, copyright statements and errata lists—each possessed of particular conventions, each with their own lively histories. To consider these paratexts—recalling them from the margins, letting them take centre stage—is to be reminded that no book is the sole work of the author whose name appears on the cover; rather, every book is the sum of a series of collaborations. It is to be reminded, also, that not everything is intended for us, the readers. There are sections that are solely directed at others—binders, librarians, lawyers—parts of the book that, if they are working well, are working discreetly, like a theatrical prompt, whispering out of the audience's ear-shot Book Parts is a bold and imaginative intervention in the fast growing field of book history: it pulls the book apart. Over twenty-two chapters, Book Parts tells the story of the components of the book: from title pages to endleaves; from dust jackets to indexes—and just about everything in between. Book Parts covers a broad historical range that runs from the pre-print era to the digital, bringing together the expertise of some of the most exciting scholars working on book history today in order to shine a new light on these elements hiding in plain sight in the books we all read.

Date	2019-06-27
Language	en
Library Catalog	Google Books
Extra	Google-Books-ID: 9CieDwAAQBAJ
Publisher	Oxford University Press

ISBN	978-0-19-254053-9
# of Pages	351
Date Added	8/4/2022, 5:22:20 AM
Modified	8/4/2022, 5:22:20 AM

- ## Tags:

  - Art / History / General
  - Design / Book
  - Literary Criticism / General

Attachments

  - Google Books Link

- # BookGPT: A Comprehensive, Step-by-Step Guide to Crafting Your Novel with Artificial Intelligence

Item Type	Book
Author	Adam Chatman IV
Abstract	Discover the groundbreaking method that is transforming the world of novel writing! Introducing the ultimate guide to harnessing the power of artificial intelligence to bring your stories to life.In this comprehensive, easy-to-use guide, you'll learn step-by-step how to utilize the groundbreaking capabilities of

ChatGPT to revolutionize your writing process. With clear instructions and practical examples, this book will help you unlock your full potential as a writer, streamline your workflow, and turn your ideas into polished, captivating manuscripts.You'll discover:How to leverage ChatGPT for every stage of your writing journey, from brainstorming to editing and everything in between.Detailed techniques to enhance your storytelling, create compelling characters, and craft immersive worlds that will captivate your readers.A foolproof method to maintain your unique voice and style while collaborating with artificial intelligence.In-depth explanations of how to make ChatGPT remember crucial story details, ensuring a consistent and engaging narrative.Time-saving hacks and expert tips to get the most out of ChatGPT, so you can focus on what matters most—your story.As a bonus, the guide includes:An extensive appendix featuring genre-specific advice and tailored prompts, making it easy to customize your writing experience no matter what kind of story you're telling.A comprehensive list of prompts and examples to help you unleash the full potential of ChatGPT in every aspect of your writing.Don't miss out on the opportunity to transform your writing process and elevate your storytelling to new heights. Join the ranks of countless authors who have already harnessed the power of ChatGPT and experience the future of novel writing today. Get your copy of "BookGPT" and begin your journey toward AI-assisted authorship now!

Language	English
Short Title	BookGPT
Library Catalog	Amazon
# of Pages	90
Date Added	6/5/2023, 4:22:13 AM
Modified	6/5/2023, 4:22:13 AM

- ## Attachments

  - ○ Amazon.com Link

- # BookGPT: A General Framework for Book Recommendation Empowered by Large Language Model

Item Type	Preprint
Author	Aakas Zhiyuli
Author	Yanfang Chen
Author	Xuan Zhang
Author	Xun Liang

Abstract With the continuous development and change exhibited by large language model (LLM) technology, represented by generative pretrained transformers (GPTs), many classic scenarios in various fields have re-emerged with new opportunities. This paper takes ChatGPT as the modeling object, incorporates LLM technology into the typical book resource understanding and recommendation scenario for the first time, and puts it into practice. By building a ChatGPT-like book recommendation system (BookGPT) framework based on ChatGPT, this paper attempts to apply ChatGPT to recommendation modeling for three typical tasks, book rating recommendation, user rating recommendation, and book summary recommendation, and explores the feasibility of LLM technology in book recommendation scenarios. At the same time, based on different evaluation schemes for book recommendation tasks and the existing classic recommendation models, this paper discusses the advantages and disadvantages of the BookGPT in book recommendation scenarios and analyzes the opportunities and improvement directions for subsequent LLMs in these scenarios.

Date	2023-05-24
Short Title	BookGPT

- # Tags:

  - ○ Computer Science - Computation and Language
  - ○ Computer Science - Information Retrieval

Notes:

  - ○ Comment: Under Review

Attachments

- arXiv Fulltext PDF
  - arXiv.org Snapshot

Books by AI (GPT-3, GPT-3.5, ChatGPT)

Item Type	Blog Post
Abstract	"It's incredible... GPT-3 beats me at writing my own novel." — Eric Silberstein (August 2021) As of 2022, there are now hundreds* of books co-authored with language models like GPT-3, using several different platforms. Usually, authors do not disclose the fact that they have used AI to co-author their book. This is a list of [...]
Date	2021-07-11T07:01:36+08:00
Language	en-AU
URL	https://lifearchitect.ai/books-by-ai/
Accessed	2/13/2023, 3:48:08 PM
Extra	WFZ Good list includes several I had not seen before.
Blog Title	Dr Alan D. Thompson – Life Architect
Date Added	2/13/2023, 3:48:07 PM
Modified	2/13/2023, 3:50:49 PM

- **Attachments**

BOOKSUM: A Collection of Datasets for Long-form Narrative Summarization

Item Type	Journal Article
Author	Wojciech Kry´scí
Author	Kry´scí Nski
Author	Nazneen Rajani
Author	Divyansh Agarwal
Author	Caiming Xiong
Author	Dragomir Radev
Author	Salesforce Research
Abstract	The majority of available text summarization datasets include short-form source documents that lack long-range causal and temporal dependencies , and often contain strong layout and stylistic biases. While relevant, such datasets will offer limited challenges for future generations of text summarization systems. We address these issues by introducing BOOK-SUM, a collection of datasets for long-form narrative summarization. Our dataset covers source documents from the literature domain, such as novels, plays and stories, and includes highly abstractive, human written summaries on three levels of granularity of increasing difficulty: paragraph-, chapter-, and book-level. The domain and structure of our dataset poses a unique set of challenges for summarization systems, which include: processing very long documents, non-trivial causal and temporal dependencies , and rich discourse structures. To facilitate future work, we trained and evaluated

multiple extractive and abstractive summariza-tion models as baselines for our dataset.

URL	http://github.com/
Accessed	9/23/2021, 8:00:00 PM
Extra	arXiv: 2105.08209v1
Date Added	9/24/2021, 5:01:41 AM
Modified	12/4/2022, 11:55:08 PM

- **Tags:**

  - #nosource
  - ⊖ No DOI found

- # Building A Virtual Machine inside ChatGPT

Item Type	Web Page
Author	Jonas DeGrave
Abstract	Unless you have been living under a rock, you have heard of this new ChatGPT assistant made by OpenAI. Did you know, that you can run a whole virtual machine inside of ChatGPT?
Date	2022-12-03T20:14:12.000Z

Language	en
URL	https://www.engraved.blog/building-a-virtual-machine-inside/
Accessed	12/5/2022, 9:21:54 AM
Extra	WFZ Money quote is this prompt. ```I want you to act as a Linux terminal. I will type commands and you will reply with what the terminal should show. I want you to only reply with the terminal output inside one unique code block, and nothing else. Do no write explanations. Do not type commands unless I instruct you to do so. When I need to tell you something in English I will do so by putting text inside curly brackets {like this}. My first command is pwd.```
Website Title	Engraved
Date Added	12/5/2022, 9:21:54 AM
Modified	12/5/2022, 11:29:37 PM

- ## **Attachments**

  - ○ Snapshot

- # BuzzFeed Stock (BZFD) Triples on Plans to Embrace OpenAI for Content - Bloomberg

Item Type	Web Page
Author	Alicia Diaz

Date	2023-01-26
URL	https://www.bloomberg.com/news/articles/2023-01-26/buzzfeed-bzfd-triples-on-plans-to-embrace-openai-for-content#xj4y7vzkg
Accessed	1/28/2023, 11:47:50 AM
Extra	WFZ The stock market loves the idea of media companies eliminating entry-level journalists to lower costs . It's pretty easy to see that the market will also love it when the big book publishing companies embrace AI. Nuclear winter for editorial assistants. Genre journeymen. Copy editors. Cover designers. And let's not forget the C suite. How hard can it be to commoditize a conglomerate CEO?
Website Title	Bloomberg
Date Added	1/28/2023, 11:47:50 AM
Modified	1/29/2023, 2:11:55 PM

- ## Attachments

  - BuzzFeed Stock (BZFD) Triples on Plans to Embrace OpenAI for Content - Bloomberg

- # ByGPT5: End-to-End Style-conditioned Poetry Generation with Token-free Language Models

Item Type	Conference Paper

Author	Jonas Belouadi
Author	Steffen Eger

Abstract	State-of-the-art poetry generation systems are often complex. They either consist of task-specific model pipelines, incorporate prior knowledge in the form of manually created constraints, or both. In contrast, end-to-end models would not suffer from the overhead of having to model prior knowledge and could learn the nuances of poetry from data alone, reducing the degree of human supervision required. In this work, we investigate end-to-end poetry generation conditioned on styles such as rhyme, meter, and alliteration. We identify and address lack of training data and mismatching tokenization algorithms as possible limitations of past attempts. In particular, we successfully pre-train ByGPT5, a new token-free decoder-only language model, and fine-tune it on a large custom corpus of English and German quatrains annotated with our styles. We show that ByGPT5 outperforms other models such as mT5, ByT5, GPT-2 and ChatGPT, while also being more parameter efficient and performing favorably compared to humans. In addition, we analyze its runtime performance and demonstrate that it is not prone to memorization. We make our code, models, and datasets publicly available.

Date	2023-07
Short Title	ByGPT5
Library Catalog	ACLWeb
URL	https://aclanthology.org/2023.acl-long.406
Accessed	7/13/2023, 2:26:23 AM
Extra	WFZ A lot of custom elements to the pipeline.
Place	Toronto, Canada
Publisher	Association for Computational Linguistics

Pages	7364–7381
Proceedings Title	Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)
Conference Name	ACL 2023
Date Added	7/13/2023, 2:26:23 AM
Modified	7/16/2023, 6:35:26 AM

- **Attachments**

  - Full Text PDF

- # CAMEL: Communicative Agents for "Mind" Exploration of Large Scale Language Model Society

Item Type	Preprint
Author	Guohao Li
Author	Hasan Abed Al Kader Hammoud
Author	Hani Itani
Author	Dmitrii Khizbullin
Author	Bernard Ghanem

Abstract	The rapid advancement of conversational and chat-based language models has led to remarkable progress in complex task-solving. However, their success heavily relies on human input to guide the conversation, which can be challenging and time-consuming. This paper explores the potential of building scalable techniques to facilitate autonomous cooperation among communicative agents and provide insight into their "cognitive" processes. To address the challenges of achieving autonomous cooperation, we propose a novel communicative agent framework named role-playing. Our approach involves using inception prompting to guide chat agents toward task completion while maintaining consistency with human intentions. We showcase how role-playing can be used to generate conversational data for studying the behaviors and capabilities of chat agents, providing a valuable resource for investigating conversational language models. Our contributions include introducing a novel communicative agent framework, offering a scalable approach for studying the cooperative behaviors and capabilities of multi-agent systems, and open-sourcing our library to support research on communicative agents and beyond. The GitHub repository of this project is made publicly available on: https://github.com/lightaime/camel.
Date	2023-03-30
Short Title	CAMEL
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2303.17760
Accessed	4/23/2023, 11:37:09 PM
Extra	arXiv:2303.17760 [cs]
DOI	10.48550/arXiv.2303.17760
Repository	arXiv
Archive ID	arXiv:2303.17760

Date Added	4/23/2023, 11:37:10 PM
Modified	4/23/2023, 11:37:13 PM

- **Tags:**

  - Computer Science - Computation and Language
  - Computer Science - Machine Learning
  - Computer Science - Artificial Intelligence
  - Computer Science - Computers and Society
  - Computer Science - Multiagent Systems

Attachments

  - arXiv Fulltext PDF
  - arXiv.org Snapshot

- # Can A.I. Invent? - The New York Times

Item Type	Web Page
URL	https://www.nytimes.com/2023/07/15/technology/ai-inventor-patents.html?unlocked_article_code=5hz6Q6gUjuFQE4K1Q03FkJoGr1lDfuOnNeB7ZrNs YF9ZuDfD6LtveHXh3eHTs-rK3a20TclU0SpjMz61grcMmng48pvYO6YAtErjsowl-6OG
Accessed	7/15/2023, 6:16:54 PM
Extra	WFZ Patent law & copyright are very differen but still worth a read.

Date Added	7/15/2023, 6:16:54 PM
Modified	7/16/2023, 6:31:57 AM

- **Attachments**
  - ○ Can A.I. Invent? - The New York Times

Can AI write good novels?

Item Type	Web Page
Author	Josh Dzieza
Abstract	Authors are getting a hand from machine learning tools — and some of them think it's the future of writing.
Date	2022-07-20T14:00:00+00:00
URL	https://www.theverge.com/c/23194235/ai-fiction-writing-amazon-kindle-sudowrite-jasper
Accessed	7/21/2022, 11:16:04 AM
Extra	WFZ Good article that shows real authors are already actively using AI as creative partners in published, profitable novels. The particular example is not super impressive -- the author compares her output of genre to "potato chips" -- but it has roughly doubled her output per year. Interesting discussions of how readers may perceive this as "cheating" and organizing to improve "truth in labeling". Hideous, difficult to read page-framed format in The Verge is a minus. HT Carl Zimmerman for the share.

Website Title	The Verge
Date Added	7/21/2022, 11:16:04 AM
Modified	7/21/2022, 7:06:46 PM

- **Notes:**

  - -

Attachments

  - Snapshot

Can GPT-3 Explain My Past and Tell My Future?

Item Type	Web Page
Author	Dan Shipper
Abstract	I loaded journal entries from the past 10 years into GPT-3—and started asking it questions
Date	2023-01-20
URL	https://every.to/chain-of-thought/can-gpt-3-explain-my-past-and-tell-me-my-future

Accessed	1/20/2023, 7:00:50 PM
Date Added	1/20/2023, 7:00:50 PM
Modified	1/20/2023, 7:00:55 PM

- **Attachments**

  - Snapshot

Can GPT-3 write an academic paper on itself, with minimal human input?

Item Type	Manuscript
Author	Gpt Generative Pretrained Transformer
Author	Almira Osmanovic Thunström
Author	Steinn Steingrimsson
Abstract	GPT-3 is a powerful artificial intelligence system that can generate text. In this paper, we explore GPT-3's ability to write about itself. We find that GPT-3 can generate clear and concise descriptions of its own capabilities and features. This is a significant advance over previous systems, which have often struggled to produce coherent text about themselves. We believe that the benefits of letting GPT-3 write about itself outweigh the risks. However, we recommend that any such writing be closely monitored by researchers in order to mitigate any potential negative consequences.
Date	2022-06

- **Notes:**

  ○ working paper or preprint

Attachments

  ○ HAL PDF Full Text

- # Can language models handle recursively nested grammatical structures? A case study on comparing models and humans

Item Type	Journal Article
Author	Andrew Lampinen

Abstract   How should we compare the capabilities of language models and humans? Here, I consider a case study: processing of recursively nested grammatical structures. Prior work has suggested that language models cannot handle these structures as reliably as humans can. However, the humans were provided with instructions and training before being evaluated, while the language models were evaluated zero-shot. I therefore attempt to more closely match the evaluation paradigms by providing language models with few-shot prompts. A simple prompt, which contains substantially less content than the human training, allows large language models to consistently outperform the human results. The same prompt even allows extrapolation to more deeply nested conditions than have been tested in humans. Further, a reanalysis of the prior human experiments suggests that the humans may not perform above chance at the difficult structures initially. These results suggest that large language models can in fact process recursively nested grammatical structures comparably to humans. This case study highlights how discrepancies in the quantity of experiment-specific context can confound comparisons of language models and humans. I use this case study to reflect on the broader challenge of comparing human and model capabilities—and to suggest that there is an important difference between evaluating cognitive models of a specific phenomenon and evaluating broadly-trained models.

Language   en

Library Catalog   Zotero

URL   https://arxiv.org/abs/2210.15303

Pages   20

Date Added   11/4/2022, 8:49:29 PM

Modified   11/4/2022, 8:51:50 PM

- # Tags:

  - ⊖ No DOI found

Attachments

- _o Lampinen - Can language models handle recursively nested gram.pdf

• CCC | Copyright licensing, content & software solutions

Item Type	Web Page
URL	https://www.copyright.com
Accessed	10/8/2022, 4:00:29 PM
Date Added	10/8/2022, 4:00:29 PM
Modified	10/8/2022, 6:58:10 PM

• Attachments

- _o CCC | Copyright licensing, content & software solutions

• CEM: Commonsense-Aware Empathetic Response Generation

Item Type	Journal Article

Author	Sahand Sabour
Author	Chujie Zheng
Author	Minlie Huang

Abstract A key trait of daily conversations between individuals is the ability to express empathy towards others, and exploring ways to implement empathy is a crucial step towards human-like dialogue systems. Previous approaches on this topic mainly focus on detecting and utilizing the user's emotion for generating empathetic responses. However, since empathy includes both aspects of affection and cognition, we argue that in addition to identifying the user's emotion, cognitive understanding of the user's situation should also be considered. To this end, we propose a novel approach for empathetic response generation, which leverages commonsense to draw more information about the user's situation and uses this additional information to further enhance the empathy expression in generated responses. We evaluate our approach on EMPATHETICDIALOGUES, which is a widely-used benchmark dataset for empathetic response generation. Empirical results demonstrate that our approach outperforms the baseline models in both automatic and human evaluations and can generate more informative and empathetic responses. Our code is available at https://github.com/Sahandfer/CEM.

Date	2022-06-28
Language	en
Short Title	CEM
Library Catalog	DOI.org (Crossref)
URL	https://ojs.aaai.org/index.php/AAAI/article/view/21373
Accessed	8/9/2022, 11:11:16 PM
Volume	36

- ## **Attachments**

    - ○ Sabour et al. - 2022 - CEM Commonsense-Aware Empathetic Response Generat.pdf

- # Character extraction and character type identification from summarised story plots

Item Type	Journal Article
Author	Vardhini Srinivasan
Author	Aurelia Power

- **Tags:**

  - #nosource

- ChatGPT cites the most-cited articles and journals, relying solely on Google Scholar's citation counts. As a result, AI may amplify the Matthew Effect in environmental science

Item Type	Preprint

Author	Eduard Petiska

| **Abstract** | ChatGPT (GPT) has become one of the most talked-about innovations in recent years, with over 100 million users worldwide. However, there is still limited knowledge about the sources of information GPT utilizes. As a result, we carried out a study focusing on the sources of information within the field of environmental science. In our study, we asked GPT to identify the ten most significant subdisciplines within the field of environmental science. We then asked it to compose a scientific review article on each subdiscipline, including 25 references. We proceeded to analyze these references, focusing on factors such as the number of citations, publication date, and the journal in which the work was published. Our findings indicate that GPT tends to cite highly-cited publications in environmental science, with a median citation count of 1184.5. It also exhibits a preference for older publications, with a median publication year of 2010, and predominantly refers to well-respected journals in the field, with Nature being the most cited journal by GPT. Interestingly, our findings suggest that GPT seems to exclusively rely on citation count data from Google Scholar for the works it cites, rather than utilizing citation information from other scientific databases such as Web of Science or Scopus. In conclusion, our study suggests that Google Scholar citations play a significant role as a predictor for mentioning a study in GPT-generated content. This finding reinforces the dominance of Google Scholar among scientific databases and perpetuates the Matthew Effect in science, where the rich get richer in terms of citations. With many scholars already utilizing GPT for literature review purposes, we can anticipate further disparities and an expanding gap between lesser-cited and highly-cited publications. |

- ## Tags:

  - Computer Science - Artificial Intelligence
  - Computer Science - Digital Libraries

Attachments

  - arXiv Fulltext PDF
  - arXiv.org Snapshot

- # ChatGPT is fun, but it is not funny! Humor is still challenging Large Language Models

Item Type	Preprint
Author	Sophie Jentzsch
Author	Kristian Kersting

Abstract   Humor is a central aspect of human communication that has not been solved for artificial agents so far. Large language models (LLMs) are increasingly able to capture implicit and contextual information. Especially, OpenAI's ChatGPT recently gained immense public attention. The GPT3-based model almost seems to communicate on a human level and can even tell jokes. Humor is an essential component of human communication. But is ChatGPT really funny? We put ChatGPT's sense of humor to the test. In a series of exploratory experiments around jokes, i.e., generation, explanation, and detection, we seek to understand ChatGPT's capability to grasp and reproduce human humor. Since the model itself is not accessible, we applied prompt-based experiments. Our empirical evidence indicates that jokes are not hard-coded but mostly also not newly generated by the model. Over 90% of 1008 generated jokes were the same 25 Jokes. The system accurately explains valid jokes but also comes up with fictional explanations for invalid jokes. Joke-typical characteristics can mislead ChatGPT in the classification of jokes. ChatGPT has not solved computational humor yet but it can be a big leap toward "funny" machines.

Date   2023-06-07

Library Catalog   arXiv.org

URL   http://arxiv.org/abs/2306.04563

Accessed   6/9/2023, 1:06:01 PM

Extra   WFZ I believe that writing humorous long-form--Wodehouse, Leacock, Sharpe--is the rarest of writing skills.

DOI   10.48550/arXiv.2306.04563

Repository   arXiv

Archive ID   arXiv:2306.04563

Date Added   6/9/2023, 1:06:01 PM

Modified	6/9/2023, 1:07:51 PM

- # Tags:

  - Computer Science - Computation and Language
  - Computer Science - Machine Learning
  - Computer Science - Artificial Intelligence
  - Computer Science - Human-Computer Interaction

Attachments

  - **arXiv Fulltext PDF**
  - **arXiv.org Snapshot**

- # ChatGPT launches boom in AI-written e-books on Amazon

Item Type	Newspaper Article
Author	Greg Bensinger
Abstract	Until recently, Brett Schickler never imagined he could be a published author, though he had dreamed about it. But after learning about the ChatGPT artificial intelligence program, Schickler figured an opportunity had landed in his lap.
Date	2023-02-21T20:43:00Z
Language	en

Library Catalog	www.reuters.com
URL	https://www.reuters.com/technology/chatgpt-launches-boom-ai-written-e-books-amazon-2023-02-21/
Accessed	2/22/2023, 9:04:43 AM
Extra	WFZ 200 books in print with author = ChatGPT as of February. None of them are described as being of high quality.
Section	Technology
Publication	Reuters
Date Added	2/22/2023, 9:04:43 AM
Modified	2/22/2023, 9:13:06 AM

- # Attachments

  - Snapshot

- # ChatGPT Will End High-School English - The Atlantic

Item Type	Magazine Article
Author	Daniel Herman
Abstract	I've been teaching English for 12 years, and I'm astounded by what ChatGPT can produce.

Date	2022-12-09
Language	English
URL	https://www.theatlantic.com/technology/archive/2022/12/openai-chatgpt-writing-high-school-english-essay/672412/?utm_source=apple_news
Accessed	12/9/2022, 7:33:39 PM
Extra	WFZ The end of high-school English will probably have minimal impact on the world of books, for the simple reason that people who write books don't learn it in high school English, and people who love books don't learn it there either. The supply of authors and readers will be unimpaired. But there will be an alarming and sad increase in the number of functoinal illiterates in the world *outside* of Booklandia--and that can't be good for us.
Publication	The Atlantic
Date Added	12/9/2022, 7:33:39 PM
Modified	12/15/2022, 8:52:53 AM

- # Tags:

  - ᵒ AI - impact on education
  - ᵒ Education

Attachments

  - ᵒ ChatGPT Will End High-School English - The Atlantic

- # China bans deepfakes created without permission or for evil

Item Type	Web Page
Author	Simon Sharwood
Abstract	'Deep synthesis service providers' otherwise free to create AI-generated humans in line with socialist values
Date	2022-12-12
Language	en
URL	https://www.theregister.com/2022/12/12/china_deep_synthesis_deepfake_regulation/
Accessed	12/13/2022, 3:19:03 AM
Date Added	12/13/2022, 3:19:03 AM
Modified	12/15/2022, 8:47:49 AM

- # Tags:

  - Safety
  - Deepfake
  - Watermarking

Notes:

- China's Cyberspace Administration has issued guidelines on how to do deepfakes the right way. Deepfakes use artificial intelligence to create realistic depictions – usually videos – of humans saying and/or doing things they didn't say and/or do. They're controversial outside China for their potential to mislead audiences and create trouble for the people depicted. Beijing clearly also has worries about the technique as the Cyberspace Administration (CAC) has issued regulations that prohibit their creation without the subject's permission, or to depict or utter anything that could be considered as counter to the national interest. Anything counter to socialist values falls under that description, as does any form of "Illegal and harmful information" or using AI-generated humans in an attempt to deceive or slander.

Attachments

- Snapshot

ChinAI #207: Reactions to ChatGPT

Item Type	Blog Post
Author	Jeffrey Ding
Abstract	Testing OpenAI's Safeguards on Questions Related to U.S.-China Relations

Date	2022-12-19
Short Title	ChinAI #207
URL	https://chinai.substack.com/p/chinai-207-reactions-to-chatgpt?publication_id=2660&isFreemail=true
Accessed	12/19/2022, 2:41:22 PM
Blog Title	ChinAI Newsletter
Website Type	Substack newsletter
Date Added	12/19/2022, 2:41:22 PM
Modified	12/19/2022, 2:41:22 PM

- # Notes:

  - thought it would be interesting to translate some reactions to this AI milestone from Zhihu, a Chinese Quora-like forum where scientists and experts often weigh in on hot topics. This thread (link to original Chinese) on ChatGPT

Attachments

- Snapshot

- # Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data

Item Type	Conference Paper
Author	Emily M. Bender
Author	Alexander Koller
Abstract	The success of the large neural language models on many NLP tasks is exciting. However, we find that these successes sometimes lead to hype in which these models are being described as "understanding" language or capturing "meaning". In this position paper, we argue that a system trained only on form has a priori no way to learn meaning. In keeping with the ACL 2020 theme of "Taking Stock of Where We've Been and Where We're Going", we argue that a clear understanding of the distinction between form and meaning will help guide the field towards better science around natural language understanding.
Date	2020
Language	en
Short Title	Climbing towards NLU
Library Catalog	Semantic Scholar
URL	https://www.aclweb.org/anthology/2020.acl-main.463
Accessed	3/3/2023, 12:10:20 AM
Extra	WFZ Important article stating the case that LLMs don't have understanding.

- # Notes:

  - [TLDR] It is argued that a system trained only on form has a priori no way to learn meaning, and a clear understanding of the distinction between form and meaning will help guide the field towards better science around natural language understanding.

Attachments

  - Full Text PDF
  - Semantic Scholar Link

Clustering by Passing Messages Between Data Points

Item Type	Journal Article
Author	Brendan J. Frey
Author	Delbert Dueck
Abstract	Clustering data by identifying a subset of representative examples is important for processing sensory signals and detecting patterns in data. Such "exemplars" can be found by randomly choosing an initial subset of data points and then iteratively refining it, but this works well only if that initial choice is close to a good solution. We devised a method called "affinity propagation," which takes as input measures of similarity between pairs of data points. Real-valued messages are exchanged between data points until a high-quality set of exemplars and corresponding clusters gradually emerges. We used affinity propagation to cluster images of faces, detect genes in microarray data, identify representative sentences in this manuscript, and identify cities that are efficiently accessed by airline travel. Affinity propagation found clusters with much lower error than other methods, and it did so in less than one-hundredth the amount of time.
Date	2007-02-16
Language	en
Library Catalog	DOI.org (Crossref)
URL	https://www.science.org/doi/10.1126/science.1136800
Accessed	8/14/2022, 11:00:18 PM
Volume	315

- **Tags:**

  - #nosource

- # CNET Is Experimenting With an AI Assist. Here's Why

Item Type	Web Page
Author	Connie Guglielmo
Abstract	For over two decades, CNET has built our reputation testing new technologies and separating the hype from reality.
Language	en

URL	https://www.cnet.com/tech/cnet-is-experimenting-with-an-ai-assist-heres-why/
Accessed	1/12/2023, 10:46:03 PM
Website Title	CNET
Date Added	1/12/2023, 10:46:03 PM
Modified	1/12/2023, 10:46:06 PM

- **Attachments**

  - Snapshot

- # CNET Is Quietly Publishing Entire Articles Generated By AI

Item Type	Web Page
Abstract	The popular tech site has employed the use of AI for its financial explainer articles under the byline of "CNET Money Staff."
URL	https://futurism.com/the-byte/cnet-publishing-articles-by-ai
Accessed	1/12/2023, 12:03:07 AM
Extra	WFZ CNET has quietly begun using GPT-like generative AI in place of reporters.
Website Title	Futurism

Date Added	1/12/2023, 12:03:07 AM
Modified	1/12/2023, 12:04:28 AM

- **Attachments**

  - Snapshot

- # CNET's AI Journalist Appears to Have Committed Extensive Plagiarism

Item Type	Web Page
Abstract	CNET's AI-generated articles appear to show deep structural similarities, amounting to plagiarism, with previously published work elsewhere.
URL	https://futurism.com/cnet-ai-plagiarism
Accessed	1/23/2023, 5:33:07 PM
Website Title	Futurism
Date Added	1/23/2023, 5:33:09 PM
Modified	1/23/2023, 5:33:09 PM

- # CNET's Article-Writing AI Is Already Publishing Very Dumb Errors

Item Type	Web Page
Abstract	The AI bot that CNET is using to churn out personal finance articles has a relatable problem: it keeps making dumb mistakes.
URL	https://futurism.com/cnet-ai-errors
Accessed	1/17/2023, 5:43:11 PM
Extra	WFZ You have to understand what generative AI is good at, and what it's not.
Website Title	Futurism
Date Added	1/17/2023, 5:43:11 PM
Modified	1/17/2023, 6:22:55 PM

- # COLLIE: Systematic Construction of Constrained Text Generation Tasks

Item Type	Web Page
Abstract	COLLIE: Systematic Construction of Constrained Text Generation Tasks
Language	en
Short Title	COLLIE
URL	https://collie-benchmark.github.io/?utm_source=tldrai

Accessed	7/19/2023, 11:04:51 AM
Date Added	7/19/2023, 11:04:51 AM
Modified	7/19/2023, 11:04:51 AM

- # Attachments

  - Snapshot

- # Colour & Shape: Using Computer Vision to Explore the Science Museum Group Collection

Item Type	Web Page
Author	Cath Sleeman
Abstract	Photographs allow us study the form of objects—their shape, colour and texture.
Date	2020-10-15T10:33:43.770Z
Language	en
Short Title	Colour & Shape
URL	https://lab.sciencemuseum.org.uk/colour-shape-using-computer-vision-to-explore-the-science-museum-c4b4f1cbd72c
Accessed	8/23/2022, 6:24:26 PM

Extra	WFZ Study of objects in science museum shows that there has been a global trend for artifacts to be less colorful in the last 20 years. Would be interesting to reproduce this with book covers.
Website Title	Medium
Date Added	8/23/2022, 6:24:26 PM
Modified	8/23/2022, 6:39:34 PM

- ## **Attachments**

  - ○ Snapshot

- # Combining keyphrase extraction and lexical diversity to characterize ideas in publication titles

Item Type	Web Page
Author	James Powell
Author	Martin Klein
Author	Lyudmila Balakireva
Abstract	Beyond bibliometrics, there is interest in characterizing the evolution of the number of ideas in scientific papers. A common approach for investigating this involves analyzing the titles of publications to detect vocabulary changes over time. With the notion that phrases, or more specifically keyphrases, represent concepts, lexical diversity metrics are applied to phrased versions of the titles.

Thus changes in lexical diversity are treated as indicators of shifts, and possibly expansion, of research. Therefore, optimizing detection of keyphrases is an important aspect of this process. Rather than just one, we propose to use multiple phrase detection models with the goal to produce a more comprehensive set of keyphrases from the source corpora. Another potential advantage to this approach is that the union and difference of these sets may provide automated techniques for identifying and omitting non-specific phrases. We compare the performance of several phrase detection models, analyze the keyphrase sets output of each, and calculate lexical diversity of corpora variants incorporating keyphrases from each model, using four common lexical diversity metrics.

Date	2022
Language	en
URL	https://www.semanticscholar.org/paper/Combining-keyphrase-extraction-and-lexical-to-ideas-Powell-Klein/9a1bf344de7096fc711a4aef8c1a438d83db8d8f
Accessed	9/1/2022, 10:23:30 AM
Date Added	9/1/2022, 10:23:30 AM
Modified	12/4/2022, 11:55:44 PM

- **Tags:**

  - #nosource

- Combining keyphrase extraction and lexical diversity to characterize ideas in publication titles | Semantic Scholar

Item Type	Web Page
URL	https://www.semanticscholar.org/paper/Combining-keyphrase-extraction-and-lexical-to-id-Klein/9a1bf344de7096fc711a4aef8c1a438d83db8d8f?utm_source=alert_email&utm_cont 0-7&utm_medium=7458942
Accessed	9/1/2022, 10:23:13 AM
Date Added	9/1/2022, 10:23:13 AM
Modified	9/1/2022, 10:23:18 AM

- ## Attachments

  - ○ Combining keyphrase extraction and lexical diversity to characterize ideas in publication titles | Semantic Scholar

- # Comparing Comprehension of a Long Text Read in Print Book and on Kindle: Where in the Text and When in the Story?

Item Type	Journal Article
Author	Anne Mangen
Author	Gérard Olivier

Author	Jean-Luc Velay
Abstract	Digital reading devices such as Kindle differ from paper books with respect to the kinesthetic and tactile feedback provided to the reader, but the role of these features in reading is rarely studied empirically. This experiment compares reading of a long text on Kindle DX and in print. Fifty participants (24 years old) read a 28 page (∼1 h reading time) long mystery story on Kindle or in a print pocket book and completed several tests measuring various levels of reading comprehension: engagement, recall, capacities to locate events in the text and reconstructing the plot of the story. Results showed that on most tests subjects performed identically whatever the reading medium. However, on measures related to chronology and temporality, those who had read in the print pocket book, performed better than those who had read on a Kindle. It is concluded that, basically comprehension was similar with both media, but, because kinesthetic feedback is less informative with a Kindle, readers were not as efficient to locate events in the space of the text and hence in the temporality of the story. We suggest that, to get a correct spatial representation of the text and consequently a coherent temporal organization of the story, readers would be reliant on the sensorimotor cues which are afforded by the manipulation of the book.
Date	2019
Short Title	Comparing Comprehension of a Long Text Read in Print Book and on Kindle
Library Catalog	Frontiers
URL	https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00038
Accessed	8/27/2022, 1:08:47 AM
Extra	WFZ Unsurprisingly, the pages of a printed book make it easier to remember where you were in the book than in a Kindle book. This could be addressed for providing a mechanical affordance--a sort of ratcheted slider--on the front of the device.
Volume	10

- **Notes:**

  - Results showed that on most tests subjects performed identically whatever the reading medium. However, on measures related to chronology and temporality, those who had read in the print pocket book, performed better than those who had read on a Kindle

Attachments

  - Full Text PDF

- # Comparing the latent space of generative models

Item Type	Preprint
Author	Andrea Asperti

Author	Valerio Tonelli
Abstract	Different encodings of datapoints in the latent space of latent-vector generative models may result in more or less effective and disentangled characterizations of the different explanatory factors of variation behind the data. Many works have been recently devoted to the exploration of the latent space of specific models, mostly focused on the study of how features are disentangled and of how trajectories producing desired alterations of data in the visible space can be found. In this work we address the more general problem of comparing the latent spaces of different models, looking for transformations between them. We confined the investigation to the familiar and largely investigated case of generative models for the data manifold of human faces. The surprising, preliminary result reported in this article is that (provided models have not been taught or explicitly conceived to act differently) a simple linear mapping is enough to pass from a latent space to another while preserving most of the information.
Date	2022-07-14
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2207.06812
Accessed	10/13/2022, 4:55:47 PM
Extra	arXiv:2207.06812 [cs]
DOI	10.48550/arXiv.2207.06812
Repository	arXiv
Archive ID	arXiv:2207.06812
Date Added	10/13/2022, 4:55:47 PM
Modified	10/13/2022, 4:55:47 PM

Attachments

- arXiv Fulltext PDF
- arXiv.org Snapshot

- # Complex Network Analysis of a Graphic Novel: The Case of the Bande Dessin ${{$$\backslash$'e$}}$ e Thorgal

Item Type	Journal Article
Author	Vincent Labatut
Date	2022
Short Title	Complex Network Analysis of a Graphic Novel
Library Catalog	Google Scholar

Publication	arXiv preprint arXiv:2206.10162
Date Added	8/5/2022, 8:53:34 PM
Modified	8/5/2022, 8:53:36 PM

- ## Tags:

  - ⊖ No DOI found

Attachments

  - Full Text
  - Snapshot

- # Composites and Correlations

Item Type	Blog Post
Author	Eryk Salvaggio
Abstract	Media Representations in the Age of AI
Date	2023-02-26
URL	https://cyberneticforests.substack.com/p/composites-and-correlations?utm_medium=email
Accessed	2/26/2023, 10:55:14 PM

Extra	WFZ Interesting exploration of tracing back from image generations to the training images using tools like haveibeentrained, which lets artists look up their names and see if they are in the dataset.
Blog Title	Cybernetic Forests
Website Type	Substack newsletter
Date Added	2/26/2023, 10:55:15 PM
Modified	2/28/2023, 2:55:06 AM

- ## Attachments

  - ○ Snapshot

- # Computational approaches to semantic change

Item Type	Book
Author	Nina Tahmasebi
Date	2021-02-26
URL	https://langsci-press.org/catalog/book/303
Accessed	9/11/2021, 8:00:00 PM
Extra	Series Title: Language Variation

Place	Berlin
Date Added	9/12/2021, 1:06:40 PM
Modified	12/4/2022, 11:55:07 PM

- **Tags:**

  - #nosource

- # ControlNet for QR Code

Item Type	Forum Post
Author	nhciao
Date	2023-06-05T15:04:33.000Z
URL	www.reddit.com/r/StableDiffusion/comments/141hg9x/controlnet_for_qr_code
Accessed	6/6/2023, 4:47:06 PM
Extra	WFZ All of retail, including major book distributors such as Ingram, are movin to QR codes to replace ISBN bar codes by 2027. This article shows that it is possible to make artistic QR codes with AI. A way for authors and publishers tc make a heretofore boring part of the book visually interesting.
Forum/Listserv Title	r/StableDiffusion
Post Type	Reddit Post

Date Added	6/6/2023, 4:47:06 PM
Modified	6/9/2023, 12:20:03 AM

- **Attachments**

  - ○ Reddit Post Snapshot

- # Creating a Testbed for the Evaluation of Automatically Generated Back-of-the-Book Indexes

Item Type	Book Section
Series Editor	David Hutchison
Series Editor	Takeo Kanade
Series Editor	Josef Kittler
Series Editor	Jon M. Kleinberg
Series Editor	Friedemann Mattern
Series Editor	John C. Mitchell
Series Editor	Moni Naor
Series Editor	Oscar Nierstrasz

Series Editor	C. Pandu Rangan
Series Editor	Bernhard Steffen
Series Editor	Madhu Sudan
Series Editor	Demetri Terzopoulos
Series Editor	Dough Tygar
Series Editor	Moshe Y. Vardi
Series Editor	Gerhard Weikum
Editor	Alexander Gelbukh
Author	Andras Csomai
Author	Rada F. Mihalcea

Abstract The automatic generation of back-of-the book indexes seems to be out of sight of the Information Retrieval and Natural Language Processing communities, although the increasingly large number of books available in electronic format, as well as recent advances in keyphrase extraction, should motivate an increased interest in this topic. In this paper, we describe the background relevant to the process of creating back-of-the-book indexes, namely (1) a short overview of the origin and structure of back-of-the-book indexes, and (2) the correspondence that can be established between techniques for automatic index construction and keyphrase extraction. Since the development of any automatic system requires in the first place an evaluation testbed, we describe our work in building a gold standard collection of books and indexes, and we present several metrics that can be used for the evaluation of automatically generated indexes against the gold standard. Finally, we investigate the properties of the gold standard index, such as index size, length of index entries, and upper bounds on coverage as indicated by the presence of index entries in the document.

Date 2006

Language	en
Library Catalog	DOI.org (Crossref)
URL	http://link.springer.com/10.1007/11671299_45
Accessed	8/22/2022, 6:18:03 PM
Extra	Series Title: Lecture Notes in Computer Science DOI: 10.1007/11671299_45
Volume	3878
Place	Berlin, Heidelberg
Publisher	Springer Berlin Heidelberg
ISBN	978-3-540-32205-4 978-3-540-32206-1
Pages	429-440
Book Title	Computational Linguistics and Intelligent Text Processing
Date Added	8/22/2022, 6:18:03 PM
Modified	8/22/2022, 6:18:04 PM

- # Attachments

  - Csomai and Mihalcea - 2006 - Creating a Testbed for the Evaluation of Automatic.pdf

- # CREATIVESUMM: Shared task on automatic summarization for creative writing

Item Type	Conference Paper
Author	Divyansh Agarwal
Author	Alexander R. Fabbri
Author	Simeng Han
Author	Wojciech Kryscinski
Author	Faisal Ladhak
Author	Bryan Li
Author	Kathleen McKeown
Author	Dragomir Radev
Author	Tianyi Zhang
Author	Sam Wiseman
Abstract	This paper introduces the shared task of summrizing documents in several creative domains, namely literary texts, movie scripts, and television scripts. Summarizing these creative documents requires making complex literary interpretations, as well as understanding non-trivial temporal dependencies in texts containing varied styles of plot development and narrative structure. This poses unique challenges and is yet underexplored for text summarization systems. In this shared task, we introduce four sub-tasks and their

corresponding datasets, focusing on summarizing books, movie scripts, primetime television scripts, and daytime soap opera scripts. We detail the process of curating these datasets for the task, as well as the metrics used for the evaluation of the submissions. As part of the CREATIVESUMM workshop at COLING 2022, the shared task attracted 18 submissions in total. We discuss the submissions and the baselines for each sub-task in this paper, along with directions for facilitating future work.

Date	2022-10
URL	https://aclanthology.org/2022.creativesumm-1.10
Extra	Citation Key: agarwal-etal-2022-creativesumm
Place	Gyeongju, Republic of Korea
Publisher	Association for Computational Linguistics
Pages	67–73
Proceedings Title	Proceedings of the workshop on automatic summarization for creative writing
Date Added	10/16/2022, 7:53:10 AM
Modified	12/4/2022, 11:55:50 PM

- **Tags:**

  - #nosource

- # Cyber Cognitive Warfare as an Emerging New War Domain and Its Strategies and Tactics

Item Type	Journal Article
Author	윤민우
Author	김은영

Abstract Cognitive warfare has become the crucial war domain that determines the outcome of modern wars. Joseph Nye pointed out, "in today's war, it is not whose army wins, but whose story wins," emphasizing the importance of narratives to occupy human minds and hearts. This has been repeatedly observed in the "War on Terror" in Afghanistan and Iraq–Syria and the Russia–Ukraine war in 2014 and 2022. The strategic importance of winning human cognition by the use of non-kinetic influence operation had been similarly emphasized in the propositions of fifth generation warfare and Gerasimov's suggestions. By recognizing the importance of cognitive warfare, this paper attempts to address the concept of cognitive warfare and suggest strategic and tactical principles for its practical operation and use. The concept of cognitive warfare is still ambiguous, blurring together with related concepts such as psychological warfare, information warfare, cyberwar, active measures, and Reflexive Control. Besides, the substantial conceptual differences between the U.S.–West and Russia have added more confusion. Thus, some clarification to distinguish the concept of cognitive warfare from other related terms seems necessary. Meanwhile, neither has there been any concrete proposition of how cognitive warfare works strategically and tactically. The literature and earlier reports only introduced various specific effects and techniques of cognitive operations. Nevertheless, how these effects and techniques are strategically–tactically integrated and jointly applied for a cohesive cognitive military operation is yet to be proposed. This paper responds to two such issues of cognitive warfare. In future warfare, the weight of cognitive warfare is anticipated to increase further. Hopefully, this paper will be a stepping stone to rouse interest in cognitive warfare and the development of its strategies and tactics.

Date	2022-12

Deep Learning based Super Resolution with OpenCV

Item Type	Web Page
Author	Xavier Weber
Abstract	Since one of the latest mergers, OpenCV contains a module that can upscale images via deep learning-based methods. Here's a step-by-step…
Date	2020-07-18T11:37:35.367Z
Language	en

URL	https://towardsdatascience.com/deep-learning-based-super-resolution-with-opencv-4fd736678066
Accessed	12/3/2022, 10:56:33 AM
Website Title	Medium
Date Added	12/3/2022, 10:56:33 AM
Modified	12/3/2022, 10:56:36 AM

- ## **Attachments**

  - Snapshot

- # Deep reinforcement learning from human preferences

Item Type	Conference Paper
Author	Paul F Christiano
Author	Jan Leike
Author	Tom Brown
Author	Miljan Martic
Author	Shane Legg

Author	Dario Amodei
Editor	I. Guyon
Editor	U. Von Luxburg
Editor	S. Bengio
Editor	H. Wallach
Editor	R. Fergus
Editor	S. Vishwanathan
Editor	R. Garnett
Date	2017
URL	https://proceedings.neurips.cc/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e4 Paper.pdf
Extra	WFZ Seminal article on "RLHP".
Volume	30
Publisher	Curran Associates, Inc.
Proceedings Title	Advances in neural information processing systems
Date Added	12/9/2022, 9:33:19 PM
Modified	12/15/2022, 8:49:52 AM

- # Deep Unsupervised Learning using Nonequilibrium Thermodynamics

Item Type	Preprint
Author	Jascha Sohl-Dickstein
Author	Eric A. Weiss
Author	Niru Maheswaranathan
Author	Surya Ganguli
Abstract	A central problem in machine learning involves modeling complex data-sets using highly flexible families of probability distributions in which learning, sampling, inference, and evaluation are still analytically or computationally tractable. Here, we develop an approach that simultaneously achieves both flexibility and tractability. The essential idea, inspired by non-equilibrium statistical physics, is to systematically and slowly destroy structure in a data distribution through an iterative forward diffusion process. We then learn a reverse diffusion process that restores structure in data, yielding a highly

flexible and tractable generative model of the data. This approach allows us to rapidly learn, sample from, and evaluate probabilities in deep generative models with thousands of layers or time steps, as well as to compute conditional and posterior probabilities under the learned model. We additionally release an open source reference implementation of the algorithm.

Date	2015-11-18
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/1503.03585
Accessed	10/2/2022, 12:00:02 PM
Extra	WFZ Pathbreaking paper for diffusion models that leads to Stable Diffusion.
DOI	10.48550/arXiv.1503.03585
Repository	arXiv
Archive ID	arXiv:1503.03585
Date Added	10/2/2022, 12:00:02 PM
Modified	10/2/2022, 12:01:36 PM

- ## **Tags:**

  - Computer Science - Machine Learning
  - Statistics - Machine Learning
  - Condensed Matter - Disordered Systems and Neural Networks

- o Quantitative Biology - Neurons and Cognition

Attachments

- o **arXiv Fulltext PDF**
- o **arXiv.org Snapshot**

- # Demystifying Prompts in Language Models via Perplexity Estimation

Item Type	Journal Article
Author	Hila Gonen
Author	Srini Iyer
Author	Terra Blevins
Author	Noah A Smith
Author	Luke Zettlemoyer
Abstract	Language models can be prompted to perform a wide variety of zero- and few-shot learning problems. However, performance varies significantly with the choice of prompt, and we do not yet understand why this happens or how to pick the best prompts. In this work, we analyze the factors that contribute to this variance and establish a new empirical hypothesis: the performance of a prompt is coupled with the extent to which the model is familiar with the language it contains. Over a wide range of tasks, we show that the lower the perplexity of the prompt is, the better the prompt is able to perform the task. As a result, we devise a method for creating prompts: (1) automatically extend a small seed set of manually written prompts by paraphrasing using GPT3 and backtranslation and (2) choose the lowest perplexity prompts to get significant gains in performance.

Language	en
Library Catalog	Zotero
Extra	WFZ Article justifying the use of perplexity to measure prompt quality; cf. betterprompt library in Python.
Date Added	1/11/2023, 9:07:19 PM
Modified	1/12/2023, 12:05:32 AM

- **Tags:**

  - ⊖ No DOI found

Attachments

  - Gonen et al. - Demystifying Prompts in Language Models via Perple.pdf

- # Digital comics image indexing based on deep learning

Item Type	Journal Article
Author	Nhu-Van Nguyen
Author	Christophe Rigaud

- # Attachments

  - Full Text
  - Snapshot

- # Discovering the Hidden Vocabulary of DALLE-2

Item Type	Preprint
Author	Giannis Daras
Author	Alexandros G. Dimakis
Abstract	We discover that DALLE-2 seems to have a hidden vocabulary that can be used to generate images with absurd prompts. For example, it seems that \texttt{Apoploe vesrreaitais} means birds and \texttt{Contarra ccetnxniams luryca tanniounons} (sometimes) means bugs or pests. We find that these prompts are often consistent in isolation but also sometimes in combinations. We present our black-box method to discover words that seem random but have some correspondence to visual concepts. This creates important security and interpretability challenges.
Date	2022-05-31
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2206.00169
Accessed	9/15/2022, 10:11:13 PM
Extra	WFZ important discoveries that may have implications for text generations. Are there "magic words" like Mrfplzik that means "create Mary Jane character"? Should these hidden vocabularies be designed in as features?
DOI	10.48550/arXiv.2206.00169
Repository	arXiv
Archive ID	arXiv:2206.00169
Date Added	9/15/2022, 10:11:13 PM

Modified	9/15/2022, 10:13:19 PM

- ## Tags:

  - ⊖ No DOI found
  - Computer Science - Computation and Language
  - Computer Science - Machine Learning
  - Computer Science - Computer Vision and Pattern Recognition
  - Computer Science - Cryptography and Security

Notes:

  - Comment: 6 pages, 4 figures

  - Comment: 6 pages, 4 figures

Attachments

  - arXiv Fulltext PDF

    ### Tags:

      - #duplicate_attachments

  - arXiv.org Snapshot
  - Daras and Dimakis - Discovering the Hidden Vocabulary of DALLE-2.pdf

Tags:

- #duplicate_attachments

- # DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter

Item Type	Preprint
Author	Victor Sanh
Author	Lysandre Debut
Author	Julien Chaumond
Author	Thomas Wolf
Abstract	As Transfer Learning from large-scale pre-trained models becomes more prevalent in Natural Language Processing (NLP), operating these large models in on-the-edge and/or under constrained computational training or inference budgets remains challenging. In this work, we propose a method to pre-train a smaller general-purpose language representation model, called DistilBERT, which can then be fine-tuned with good performances on a wide range of tasks like its larger counterparts. While most prior work investigated the use of distillation for building task-specific models, we leverage knowledge distillation during the pre-training phase and show that it is possible to reduce the size of a BERT model by 40%, while retaining 97% of its language understanding capabilities and being 60% faster. To leverage the inductive biases learned by larger models during pre-training, we introduce a triple loss combining language modeling, distillation and cosine-distance losses. Our smaller, faster and lighter model is cheaper to pre-train and we demonstrate its capabilities for on-device computations in a proof-of-concept experiment and a comparative on-device study.

Date	2020-02-29
Short Title	DistilBERT, a distilled version of BERT
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/1910.01108
Accessed	9/11/2022, 2:10:12 PM
Extra	arXiv:1910.01108 [cs]
DOI	10.48550/arXiv.1910.01108
Repository	arXiv
Archive ID	arXiv:1910.01108
Date Added	9/11/2022, 2:10:12 PM
Modified	9/11/2022, 2:10:16 PM

- **Tags:**

  - Computer Science - Computation and Language

Notes:

  - prior

Attachments

- o arXiv Fulltext PDF
- o arXiv.org Snapshot

- ## Does Artificial Intelligence Really Have the Potential to Create Transformative Art?

Item Type	Blog Post
Abstract	I. The Situation In 1896, the Lumiere brothers released a 50-second-long film, The Arrival of a Train at La Ciotat, and a myth was born. The audiences, it was reported, were so entranced by the new…
Date	2022-06-27T08:59:02+00:00
Language	en-US
URL	https://lithub.com/does-artificial-intelligence-really-have-the-potential-to-create-transformative-art/
Accessed	6/28/2022, 3:43:47 PM
Extra	WFZ Thoughtful essay.

Blog Title	Literary Hub
Date Added	6/28/2022, 3:43:47 PM
Modified	12/4/2022, 11:55:29 PM

- **Tags:**

  - #nosource

- # E Ink Is Awarded the World's First ePaper Module Carbon Footprint Verification

Item Type	Blog Post
Author	Michael Kozlowski
Abstract	E INK the originator, pioneer, and global commercial leader in digital paper technology, today announced that the carbon footprint of their 6.8 inch e-paper module for e-readers and their 2.9 inch e-paper module for electronic shelf labels (ESL) have been verified by the British Standards Institution (BSI), as being in compliance with the ISO 14067:2018 standard. The carbon footprint of every 6.8 inch e-paper module is 3.30 kg CO2; every 2.9 inch e-paper module is 0.59 kg CO2e. In December 2021, E Ink announced their pledge to reach 100% use of renewable energy by 2030 and net zero carbon emissions
Date	2022-08-26T03:11:29-04:00
Language	en-US

URL	https://goodereader.com/blog/e-paper/e-ink-is-awarded-the-worlds-first-epaper-module-carbon-footprint-verification
Accessed	8/28/2022, 6:52:50 PM
Blog Title	Good e-Reader
Date Added	8/28/2022, 6:52:50 PM
Modified	8/28/2022, 6:52:55 PM

- **Notes:**

  - verified by the British Standards Institution (BSI), as being in compliance with the ISO 14067:2018 standard. The carbon footprint of every 6.8 inch e-paper module is 3.30 kg CO_2;

Attachments

  - Snapshot

- ## Eccentric Dictionaries - An Experiment in AI-Enhanced Human Creativity

Item Type	Book

Author	Zimmerman , Frederick
Date	2021
URL	https://www.google.com/books/edition/Eccentric_Dictionaries/JgU5zgEACAAJ?hl=en&
Accessed	8/11/2021, 8:00:00 PM
Extra	WFZ Pioneering!
Place	Ann Arbor
Publisher	Nimble Books LLC
Date Added	8/12/2021, 3:04:56 AM
Modified	12/4/2022, 11:55:07 PM

- **Tags:**

  - #nosource

Effidit: Your AI writing assistant

Item Type	Document
Author	Shuming Shi
Author	Enbo Zhao

- **Tags:**

  - #nosource
  - Computation and Language (cs.CL)
  - FOS: Computer and information sciences

- # Emotion and Modifier in Henry Rider Haggard's Novels

Item Type	Conference Paper
Author	Salim Sazzed
Abstract	In recent years, there has been a growing scholarly interest in employing quantitative methods to analyze literary texts, as they offer unique insights, theories, and interpretations. In light of this, the current study employs quantitative analysis to examine the fiction written by the renowned British adventure novelist, Sir Henry Rider Haggard. Specifically, the study aims to investigate the affective content and prevalence of distinctive linguistic features in six of Haggard's most distinguished works. We evaluate dominant emotional states at the sentence level as well as investigate the deployment of specific linguistic features such as modifiers and deontic modals, and collocated terms. Through sentence-level emotion analysis the findings reveal a notable prevalence of "joy"-related emotions across the novels. Furthermore, the study observes that intensifiers are employed more commonly than the mitigators as modifiers and the collocated terms of modifiers exhibit high similarity across the novels. By integrating quantitative analyses with qualitative assessments, this study presents a novel perspective on the patterns of emotion and specialized grammatical features in some of Haggard's most celebrated literary works.
Date	2023-07

Library Catalog	ACLWeb
URL	
Accessed	7/13/2023, 2:27:29 AM
Place	Toronto, Canada
Publisher	Association for Computational Linguistics
Pages	11–15
Proceedings Title	Proceedings of the The 5th Workshop on Narrative Understanding
Conference Name	WNU 2023
Date Added	7/13/2023, 2:27:29 AM
Modified	7/13/2023, 2:27:29 AM

- **Attachments**

  - Full Text PDF

- # Emotion Recognition Based on the Structure of Narratives

Item Type	Journal Article

Author	Tibor Pólya
Author	István Csertő
Abstract	One important application of natural language processing (NLP) is the recognition of emotions in text. Most current emotion analyzers use a set of linguistic features such as emotion lexicons, n-grams, word embeddings, and emoticons. This study proposes a new strategy to perform emotion recognition, which is based on the homologous structure of emotions and narratives. It is argued that emotions and narratives share both a goal-based structure and an evaluation structure. The new strategy was tested in an empirical study with 117 participants who recounted two narratives about their past emotional experiences, including one positive and one negative episode. Immediately after narrating each episode, the participants reported their current affective state using the Affect Grid. The goal-based structure and evaluation structure of the narratives were analyzed with a hybrid method. First, a linguistic analysis of the texts was carried out, including tokenization, lemmatization, part-of-speech tagging, and morphological analysis. Second, an extensive set of rule-based algorithms was used to analyze the goal-based structure of, and evaluations in, the narratives. Third, the output was fed into machine learning classifiers of narrative structural features that previously proved to be effective predictors of the narrator's current affective state. This hybrid procedure yielded a high average F1 score (0.72). The results are discussed in terms of the benefits of employing narrative structure analysis in NLP-based emotion recognition.
Date	2023/1
Language	en
Library Catalog	www.mdpi.com
URL	https://www.mdpi.com/2079-9292/12/4/919
Accessed	2/13/2023, 9:41:18 PM
Rights	

- ## Tags:

  - machine learning
  - emotion lexicon
  - emotion recognition
  - narrative structure

Attachments

  - Full Text PDF

End-to-End Chinese Speaker Identification

Item Type	Conference Paper
Author	Dian Yu
Author	Ben Zhou
Author	Dong Yu
Date	2022
Library Catalog	Google Scholar
Pages	2274–2285
Proceedings Title	Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies
DOI	10.18653/v1/2022.naacl-main.165
Date Added	8/5/2022, 8:53:34 PM
Modified	8/5/2022, 8:53:38 PM

Attachments

- Full Text
- Snapshot

Entity-based SpanCopy for Abstractive Summarization to Improve the Factual Consistency

Item Type	Preprint
Author	Wen Xiao
Author	Giuseppe Carenini
Abstract	Despite the success of recent abstractive summarizers on automatic evaluation metrics, the generated summaries still present factual inconsistencies with the source document. In this paper, we focus on entity-level factual inconsistency, i.e. reducing the mismatched entities between the generated summaries and the source documents. We therefore propose a novel entity-based SpanCopy mechanism, and explore its extension with a Global Relevance component. Experiment results on four summarization datasets show that SpanCopy can effectively improve the entity-level factual consistency with essentially no change in the word-level and entity-level saliency. The code is available at https://github.com/Wendy-Xiao/Entity-based-SpanCopy
Date	2022-09-07
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2209.03479
Accessed	9/13/2022, 3:54:32 AM
Extra	arXiv:2209.03479 [cs]
DOI	10.48550/arXiv.2209.03479

Repository	arXiv
Archive ID	arXiv:2209.03479
Date Added	9/13/2022, 3:54:32 AM
Modified	9/13/2022, 3:54:32 AM

- ## Tags:

  - Computer Science - Computation and Language

Attachments

  - arXiv Fulltext PDF
  - arXiv.org Snapshot

- # Exclusive: ChatGPT owner OpenAI projects $1 billion in revenue by 2024

Item Type	Newspaper Article
Author	Jeffrey Dastin
Author	Krystal Hu
Author	Paresh Dave
Author	Paresh Dave

Abstract	The forecast represents how some are betting the AI tech will go far beyond splashy and sometimes flawed public demos.
Date	2022-12-15T15:09:29Z
Language	en
Short Title	Exclusive
Library Catalog	www.reuters.com
URL	https://www.reuters.com/business/chatgpt-owner-openai-projects-1-billion-revenue-by-2024-sources-2022-12-15/
Accessed	12/16/2022, 5:11:24 AM
Extra	WFZ More hard numbers around the large-language model business than I have seen elsewhere. Jasper marketing content app has 80,000 customers at 29USD/mo. OpenAI expects 200M USB in revenue in 2023. And I don't think they have found the big enterprise applications yet. Some good inside baseball stuff about valuations. OpenAI cap structure has some oddities--return capped at 100x.
Section	Business
Publication	Reuters
Date Added	12/16/2022, 5:11:24 AM
Modified	12/16/2022, 6:06:56 AM

- # **Attachments**

- o Snapshot

Explainable AI Course

Item Type	Web Page
URL	https://interpretable-ml-class.github.io/
Accessed	6/28/2023, 3:32:00 AM
Date Added	6/28/2023, 3:32:00 AM
Modified	6/28/2023, 3:32:00 AM

Attachments

- o Explainable AI Course

Exploratory network analysis of Marvel Universe

Item Type	Web Page
Author	Tomaz Bratanic
Abstract	Introducing the new k-nearest neighbors algorithm in the Graph Data Science library
Date	2020-10-26T17:25:24.040Z

Language	en
URL	https://towardsdatascience.com/exploratory-network-analysis-of-marvel-universe-c557f4959048
Accessed	2/7/2022, 12:30:56 AM
Extra	WFZ Some interesting reference points here. The MCU, which is worth a gazillion dollars, contains 1,105 characters in 38,875 comics.
Website Title	Medium
Date Added	2/7/2022, 12:30:56 AM
Modified	12/4/2022, 11:55:10 PM

- **Tags:**

  - #nosource

- # Exploring the Boundaries of AI: How a Human-Machine Collaboration Created a Book on Our Future

Item Type	Blog Post
Author	Dr Mark van Rijmenam

| Abstract | I just wrote a book in one week ☐☐☐ Yes, thanks to #ChatGPT—a great example of human-machine collaboration. |

Date	2022-12-12T08:50:49.575Z
Language	en
Short Title	Exploring the Boundaries of AI
URL	https://markvanrijmenam.medium.com/exploring-the-boundaries-of-ai-how-a-human-machine-collaboration-created-a-book-on-our-future-3e244af50941
Accessed	12/13/2022, 11:35:44 PM
Extra	WFZ Author did seven-day experiment creating book with ChatGPT, exploring its strengths & weaknesses. Seems to have erred on the side of deferring to ChatGPT's wording rather than heavily editing it.
Blog Title	Medium
Date Added	12/13/2022, 11:35:44 PM
Modified	12/15/2022, 8:44:24 AM

- # Tags:

  - Books and Literature
  - Writing and Writers

Attachments

- o Snapshot

- # Extraction and analysis of fictional character networks: A survey

Item Type	Journal Article
Author	Vincent Labatut
Author	Xavier Bost
Abstract	A character network is a graph extracted from a narrative in which vertices represent characters and edges correspond to interactions between them. A number of narrative-related problems can be addressed automatically through the analysis of character networks, such as summarization, classification, or role detection. Character networks are particularly relevant when considering works of fiction (e.g., novels, plays, movies, TV series), as their exploitation allows developing information retrieval and recommendation systems. However, works of fiction possess specific properties that make these tasks harder. This survey aims at presenting and organizing the scientific literature related to the extraction of character networks from works of fiction, as well as their analysis. We first describe the extraction process in a generic way and explain how its constituting steps are implemented in practice, depending on the medium of the narrative, the goal of the network analysis, and other factors. We then review the descriptive tools used to characterize character networks, with a focus on the way they are interpreted in this context. We illustrate the relevance of character networks by also providing a review of applications derived from their analysis. Finally, we identify the limitations of the existing approaches and the most promising perspectives.
Date	2019-09-01
Language	en
Short Title	Extraction and Analysis of Fictional Character Networks

Library Catalog	DOI.org (Crossref)
URL	https://dl.acm.org/doi/10.1145/3344548
Accessed	8/21/2021, 8:00:00 PM
Extra	Publisher: Association for Computing Machinery
Volume	52
Pages	1-40
Publication	ACM Computing Surveys
DOI	10.1145/3344548
Issue	5
Journal Abbr	ACM Comput. Surv.
ISSN	0360-0300, 1557-7341
Date Added	8/22/2021, 12:28:26 PM
Modified	10/19/2022, 11:01:23 PM

- **Tags:**

  - Character network
  - Graph analysis
  - Graph extraction

- o Information retrieval
- o Narrative
- o Work of fiction
- o #nosource

Attachments

- o Labatut and Bost - 2020 - Extraction and Analysis of Fictional Character Net.pdf

  ### Tags:

  - #duplicate_attachments

- o Submitted Version

  ### Tags:

  - #duplicate_attachments

Extractive is not Faithful: An Investigation of Broad Unfaithfulness Problems in Extractive Summarization

Item Type	Preprint

Author	Shiyue Zhang
Author	David Wan
Author	Mohit Bansal

Abstract   The problems of unfaithful summaries have been widely discussed under the context of abstractive summarization. Though extractive summarization is less prone to the common unfaithfulness issues of abstractive summaries, does that mean extractive is equal to faithful? Turns out that the answer is no. In this work, we define a typology with five types of broad unfaithfulness problems (including and beyond not-entailment) that can appear in extractive summaries, including incorrect coreference, incomplete coreference, incorrect discourse, incomplete discourse, as well as other misleading information. We ask humans to label these problems out of 1500 English summaries produced by 15 diverse extractive systems. We find that 33% of the summaries have at least one of the five issues. To automatically detect these problems, we find that 5 existing faithfulness evaluation metrics for summarization have poor correlations with human judgment. To remedy this, we propose a new metric, ExtEval, that is designed for detecting unfaithful extractive summaries and is shown to have the best performance. We hope our work can increase the awareness of unfaithfulness problems in extractive summarization and help future work to evaluate and resolve these issues. Our data and code are publicly available at https://github.com/ZhangShiyue/extractive_is_not_faithful

Date	2022-09-07
Short Title	Extractive is not Faithful
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2209.03549
Accessed	9/16/2022, 10:35:59 PM
Extra	WFZ Reviews existing automatic metrics for testing faithfulness of summarization and introduces a new method, ExtEval.

- ## Tags:

    - Computer Science - Computation and Language
    - Computer Science - Artificial Intelligence

Notes:

    - Comment: 19 pages

Attachments

    - arXiv Fulltext PDF
    - arXiv.org Snapshot

- # Fable unveils Showrunner AI to create South Park-like TV shows with you as the star

Item Type	Blog Post
Abstract	Fable has released Showrunner AI technology, dubbed SHOW-1, which can generate new episodes of TV shows with you as the star.
Date	2023-07-18T17:00:00+00:00
Language	en-US
URL	https://venturebeat.com/games/the-simulation-unveils-showrunner-ai-to-create-south-park-like-tv-shows-with-you-as-the-star/
Accessed	7/19/2023, 11:13:42 PM
Extra	WFZ One review says "used AI to create a mediocre episode of South Park". Already a show that does less than nothing for me.
Blog Title	VentureBeat
Date Added	7/19/2023, 11:13:41 PM
Modified	7/21/2023, 9:37:47 PM

- ## **Attachments**

  - ○ Snapshot

- # FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial

Intelligence Companies to Manage the Risks Posed by AI

Item Type	Web Page
Author	The White House
Abstract	Voluntary commitments – underscoring safety, security, and trust – mark a critical step toward developing responsible AI Biden-Harris Administration will continue to take decisive action by developing an Executive Order and pursuing bipartisan legislation to keep Americans safe Since taking office, President Biden, Vice President Harris, and the entire Biden-Harris Administration have moved with urgency…
Date	2023-07-21T09:00:00+00:00
Language	en-US
Short Title	FACT SHEET
URL	https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/
Accessed	7/21/2023, 5:23:21 AM
Extra	WFZ The censors are at it again.
Website Title	The White House
Date Added	7/21/2023, 5:23:20 AM
Modified	7/21/2023, 9:35:44 PM

- **Attachments**

  - Snapshot

- # FastComposer: Tuning-free multi-subject image generation with localized attention

Item Type	Journal Article
Author	Guangxuan Xiao
Author	Tianwei Yin
Author	William T. Freeman
Author	Frédo Durand
Author	Song Han
Date	2023
Extra	WFZ MIT
Publication	arXiv
Date Added	5/22/2023, 3:00:13 AM
Modified	5/23/2023, 8:04:46 PM

- **Tags:**

- # Faster sorting algorithms discovered using deep reinforcement learning

Item Type	Journal Article
Author	Daniel J. Mankowitz
Author	Andrea Michi
Author	Anton Zhernov
Author	Marco Gelmi
Author	Marco Selvi
Author	Cosmin Paduraru
Author	Edouard Leurent
Author	Shariq Iqbal
Author	Jean-Baptiste Lespiau
Author	Alex Ahern
Author	Thomas Köppe

Author   Kevin Millikin

Author   Stephen Gaffney

Author   Sophie Elster

Author   Jackson Broshear

Author   Chris Gamble

Author   Kieran Milan

Author   Robert Tung

Author   Minjae Hwang

Author   Taylan Cemgil

Author   Mohammadamin Barekatain

Author   Yujia Li

Author   Amol Mandhane

Author   Thomas Hubert

Author   Julian Schrittwieser

Author   Demis Hassabis

Author   Pushmeet Kohli

Author   Martin Riedmiller

Author	Oriol Vinyals
Author	David Silver

Abstract — Fundamental algorithms such as sorting or hashing are used trillions of times on any given day1. As demand for computation grows, it has become critical for these algorithms to be as performant as possible. Whereas remarkable progress has been achieved in the past2, making further improvements on the efficiency of these routines has proved challenging for both human scientists and computational approaches. Here we show how artificial intelligence can go beyond the current state of the art by discovering hitherto unknown routines. To realize this, we formulated the task of finding a better sorting routine as a single-player game. We then trained a new deep reinforcement learning agent, AlphaDev, to play this game. AlphaDev discovered small sorting algorithms from scratch that outperformed previously known human benchmarks. These algorithms have been integrated into the LLVM standard C++ sort library3. This change to this part of the sort library represents the replacement of a component with an algorithm that has been automatically discovered using reinforcement learning. We also present results in extra domains, showcasing the generality of the approach.

Date	2023-06
Language	en
Library Catalog	www.nature.com
URL	https://www.nature.com/articles/s41586-023-06004-9
Accessed	6/8/2023, 3:33:39 AM
Rights	2023 The Author(s)
Extra	Number: 7964 Publisher: Nature Publishing Group
Volume	618

- **Tags:**

  - Software
  - Computer science

Attachments

  - Full Text PDF

- # Fiction Popularity Prediction Based on Emotion Analysis

Item Type	Conference Paper
Author	Xing Wang

Author	Shouhua Zhang
Author	Ivan Smetannikov
Date	2020-10-27
Language	en
Library Catalog	DOI.org (Crossref)
URL	https://dl.acm.org/doi/10.1145/3437802.3437831
Accessed	3/11/2022, 3:47:57 PM
Place	Xiamen China
Publisher	ACM
ISBN	978-1-4503-8805-4
Pages	169-175
Proceedings Title	2020 International Conference on Control, Robotics and Intelligent System
Conference Name	CCRIS 2020: 2020 International Conference on Control, Robotics and Intelligent System
DOI	10.1145/3437802.3437831
Date Added	3/11/2022, 3:47:58 PM
Modified	12/4/2022, 11:55:11 PM

- **Tags:**

File Creation Guide

Item Type	Document
Author	Lightning Source, Inc.
Abstract	Technical information about how to prepare files for submission to POD vendor Lightning Source.
Date	2022-10-25
URL	https://www.ingramcontent.com/publishers-document/file-creation-guide
Accessed	12/20/2022, 1:58:43 AM
Extra	WFZ Most AI-to-printed-book solutions will involve the two major POD distributors/printers, Lightning Source and Kindle Direct Publishing. This is the requirements bible for LSI.
Publisher	Ingram
Date Added	12/20/2022, 1:58:43 AM
Modified	12/22/2022, 11:27:39 PM

Fine-Tuning Language Models from Human Preferences

| Item Type | Journal Article |

Author Daniel M. Ziegler

Author Nisan Stiennon

Author Jeffrey Wu

Author Tom B. Brown

Author Alec Radford

Author Dario Amodei

Author Paul Christiano

Author Geoffrey Irving

Abstract Reward learning enables the application of reinforcement learning (RL) to tasks where reward is defined by human judgment, building a model of reward by asking humans questions. Most work on reward learning has used simulated environments, but complex information about values is often expressed in natural language, and we believe reward learning for language is a key to making RL practical and safe for real-world tasks. In this paper, we build on advances in generative pretraining of language models to apply reward learning to four natural language tasks: continuing text with positive sentiment or physically descriptive language, and summarization tasks on the TL;DR and CNN/Daily Mail datasets. For stylistic continuation we achieve good results with only 5,000 comparisons evaluated by humans. For summarization, models trained with 60,000 comparisons copy whole sentences from the input but skip irrelevant preamble; this leads to reasonable ROUGE scores and very good performance according to our human labelers, but may be exploiting the fact that labelers rely on simple heuristics.

Date 2019

Library Catalog	DOI.org (Datacite)
URL	https://arxiv.org/abs/1909.08593
Accessed	10/19/2022, 9:01:51 PM
Rights	arXiv.org perpetual, non-exclusive license
Extra	Publisher: arXiv Version Number: 2
DOI	10.48550/ARXIV.1909.08593
Date Added	10/19/2022, 9:01:51 PM
Modified	12/4/2022, 11:55:50 PM

- **Tags:**

  - #nosource
  - Computation and Language (cs.CL)
  - FOS: Computer and information sciences
  - Machine Learning (cs.LG)
  - Machine Learning (stat.ML)

- # FLASK: Fine-grained Language Model Evaluation based on Alignment Skill Sets

Item Type	Preprint

Author   Seonghyeon Ye

Author   Doyoung Kim

Author   Sungdong Kim

Author   Hyeonbin Hwang

Author   Seungone Kim

Author   Yongrae Jo

Author   James Thorne

Author   Juho Kim

Author   Minjoon Seo

Abstract   Evaluation of Large Language Models (LLMs) is challenging because aligning to human values requires the composition of multiple skills and the required set of skills varies depending on the instruction. Recent studies have evaluated the performance of LLMs in two ways, (1) automatic evaluation on several independent benchmarks and (2) human or machined-based evaluation giving an overall score to the response. However, both settings are coarse-grained evaluations, not considering the nature of user instructions that require instance-wise skill composition, which limits the interpretation of the true capabilities of LLMs. In this paper, we introduce FLASK (Fine-grained Language Model Evaluation based on Alignment SKill Sets), a fine-grained evaluation protocol that can be used for both model-based and human-based evaluation which decomposes coarse-level scoring to an instance-wise skill set-level. Specifically, we define 12 fine-grained skills needed for LLMs to follow open-ended user instructions and construct an evaluation set by allocating a set of skills for each instance. Additionally, by annotating the target domains and difficulty level for each instance, FLASK provides a holistic view with a comprehensive analysis of a model's performance depending on skill, domain, and difficulty. Through using FLASK, we compare multiple open-sourced and proprietary LLMs and observe highly-correlated findings between model-based and human-based evaluations. FLASK enables developers to more accurately measure the model performance and how it can be improved by analyzing factors that make LLMs

proficient in particular skills. For practitioners, FLASK can be used to recommend suitable models for particular situations through comprehensive comparison among various LLMs. We release the evaluation data and code implementation at https://github.com/kaistAI/FLASK.

Date	2023-07-20
Short Title	FLASK
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2307.10928
Accessed	7/21/2023, 11:03:54 PM
Extra	WFZ A good way to measure models at a glance, but for specific publishing applications, you need to experiment hands-on. https://www.dropbox.com/s/icxc1tlw2gkvijd/Screenshot%202023-07-22%20at%204.56.44%20AM.png?dl=0
DOI	10.48550/arXiv.2307.10928
Repository	arXiv
Archive ID	arXiv:2307.10928
Date Added	7/21/2023, 11:03:54 PM
Modified	7/22/2023, 4:58:17 AM

- # Tags:

  - Computer Science - Computation and Language

- o Computer Science - Artificial Intelligence

Attachments

- o **arXiv Fulltext PDF**
- o **arXiv.org Snapshot**

• Forgetify: Popular Music That Nobody Listens To

Item Type	Web Page
Author	Daniel Antal
Abstract	Forgetify is an application that is "recommending" you songs that have never been played on Spotify - not even by their families, friends or foes. When you design a recommendation engine for an artist or a label, you want to avoid that their songs ever arrive to Forgetify.
Date	2020-10-24T18:00:00+02:00
Language	en-us
Short Title	Forgetify
URL	https://dataandlyrics.com/post/2020-10-24-forgetify_pop_october/
Accessed	11/7/2022, 12:56:28 AM
Extra	WFZ Explores the long tail of copyrighted music via a look at Forgetify. Some interesting lessons for those interested in the long tail of copyrighted books.

Website Title	Data & Lyrics
Date Added	11/7/2022, 12:56:28 AM
Modified	11/7/2022, 1:02:43 AM

- **Attachments**

  - Snapshot

- # From subtitles to substantial metadata: examining characteristics of named entities and their role in indexing

Item Type	Journal Article
Author	Anne-Stine Ruud Husevåg
Date	2019
Short Title	From subtitles to substantial metadata
Library Catalog	Google Scholar
Extra	Publisher: Springer

- ## Attachments

  - Full Text
  - Snapshot

- # Future Visions: A human-machine collaboration on the potential of technology

Item Type	Book
Author	Mark van Rijmenam
Author	ChatGPT- OpenAI
Date	2022-12-11

Language	English
Short Title	Future Visions
Library Catalog	Amazon
Extra	WFZ Amazon link: https://amzn.to/3V2UvQy
Publisher	The Digital Speaker
# of Pages	151
Date Added	12/13/2022, 11:56:59 PM
Modified	12/15/2022, 8:43:54 AM

- **Tags:**

  - Books and Literature
  - Writing and Writers

Attachments

  - Amazon.com Link

- # Generative Agents: Interactive Simulacra of Human Behavior

Item Type	Preprint
Author	Joon Sung Park
Author	Joseph C. O'Brien
Author	Carrie J. Cai
Author	Meredith Ringel Morris
Author	Percy Liang
Author	Michael S. Bernstein

Abstract   Believable proxies of human behavior can empower interactive applications ranging from immersive environments to rehearsal spaces for interpersonal communication to prototyping tools. In this paper, we introduce generative agents--computational software agents that simulate believable human behavior. Generative agents wake up, cook breakfast, and head to work; artists paint, while authors write; they form opinions, notice each other, and initiate conversations; they remember and reflect on days past as they plan the next day. To enable generative agents, we describe an architecture that extends a large language model to store a complete record of the agent's experiences using natural language, synthesize those memories over time into higher-level reflections, and retrieve them dynamically to plan behavior. We instantiate generative agents to populate an interactive sandbox environment inspired by The Sims, where end users can interact with a small town of twenty five agents using natural language. In an evaluation, these generative agents produce believable individual and emergent social behaviors: for example, starting with only a single user-specified notion that one agent wants to throw a Valentine's Day party, the agents autonomously spread invitations to the party over the next two days, make new acquaintances, ask each other out on dates to the party, and coordinate to show up for the party together at the right time. We demonstrate through ablation that the components of our agent architecture--observation, planning, and reflection--each contribute critically to the believability of agent behavior. By fusing large language models with computational, interactive agents, this work introduces architectural and interaction patterns for enabling believable simulations of human behavior.

Date	2023-04-06
Short Title	Generative Agents
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2304.03442
Accessed	4/10/2023, 12:45:44 PM
Extra	arXiv:2304.03442 [cs]
DOI	10.48550/arXiv.2304.03442
Repository	arXiv
Archive ID	arXiv:2304.03442
Date Added	4/10/2023, 12:45:44 PM
Modified	4/10/2023, 12:45:44 PM

- **Tags:**

    - Computer Science - Machine Learning
    - Computer Science - Artificial Intelligence
    - Computer Science - Human-Computer Interaction

Attachments

- o arXiv Fulltext PDF
- o arXiv.org Snapshot

Generative AI and intellectual property

Item Type	Web Page
Abstract	If you put all the world's knowledge into an AI model and use it to make something new, who owns that and who gets paid? This is a completely new problem that we've been arguing about for 500 years.
Date	2023-08-27
Language	en-GB
URL	https://www.ben-evans.com/benedictevans/2023/8/27/generative-ai-ad-intellectual-property
Accessed	8/28/2023, 3:47:26 PM
Website Title	Benedict Evans
Date Added	8/28/2023, 3:47:26 PM
Modified	8/28/2023, 3:47:37 PM

Attachments

- o Snapshot

- # Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations

Item Type	Journal Article
Author	Stanford Internet Observatory
Author	OpenAI
Author	and Georgetown University's Center for Security and Emerging Technology
Date	Wed, 01/11/2023 - 12:00
Language	en
Short Title	Generative Language Models and Automated Influence Operations
Library Catalog	cyber.fsi.stanford.edu
URL	https://cyber.fsi.stanford.edu/io/publication/generative-language-models-and-automated-influence-operations-emerging-threats-and
Accessed	1/12/2023, 10:12:29 PM
Date Added	1/12/2023, 10:12:29 PM
Modified	1/12/2023, 10:12:32 PM

- **Attachments**

  - ○ Snapshot

- # Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations

Item Type	Journal Article
Author	Josh A Goldstein
Author	Girish Sastry
Author	Micah Musser
Author	Renée DiResta
Author	Matthew Gentzel
Author	Katerina Sedova
Language	en
Library Catalog	Zotero
Date Added	1/12/2023, 10:12:53 PM
Modified	1/12/2023, 10:12:53 PM

- **Attachments**

  - Goldstein et al. - Generative Language Models and Automated Influence.pdf

- # Geodesic cycle length distributions in fictional character networks

Item Type	Journal Article
Author	Alex Stivala
Abstract	A geodesic cycle in a graph is a cycle with no shortcuts, so that the shortest path between any two nodes in the cycle is the path along the cycle itself. A recently published paper used random graph models to investigate the geodesic cycle length distributions of a unique set of delusional social networks, first examined in an earlier work, as well as some other publicly available social networks. Here I test the hypothesis, suggested in the former work, that fictional character networks, and in particular those from works by a single author, might have geodesic cycle length distributions which are extremely unlikely under random graph models, as the delusional social networks do. The results do not show any support for this hypothesis. In addition, the recently published work is reproduced using a method for counting geodesic cycles exactly, rather than the approximate method used originally. The substantive conclusions of that work are unchanged, but some differences in the results for particular networks are described.
Language	en
Library Catalog	Zotero
Date Added	3/24/2023, 8:37:59 PM
Modified	3/24/2023, 8:38:00 PM

- **Tags:**

  - ⊖ No DOI found

Attachments

  - Stivala - Geodesic cycle length distributions in fictional ch.pdf

- # Getty Images is suing the creators of AI art tool Stable Diffusion for scraping its content

Item Type	Web Page
Author	James Vincent
Abstract	Getty Images claims Stability AI 'unlawfully' scraped millions of images from its site. It's a significant escalation in the developing legal battles between generative AI firms and content creators.
Date	2023-01-17T10:30:16.195Z
Language	en-US
URL	https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit
Accessed	1/17/2023, 5:46:36 PM

Extra	WFZ Getty is not trying to shut Stable Diffusion down, they are looking for a licensing regime. While I am not generally very sympathetic to legal challenges to generative AI, the fact that SD frequently adds Getty Images watermarks to its generated images makes this a bit different! That's just lame.
Website Title	The Verge
Date Added	1/17/2023, 5:46:36 PM
Modified	1/17/2023, 6:22:06 PM

- # **Attachments**

  - ## Snapshot

- # Ghost Writer: Microsoft Looks to Add OpenAI's Chatbot Technology to Word, Email

Item Type	Magazine Article
Author	Aaron Holmes
Author	Kevin McLaughlin
Abstract	In a move that could change how more than a billion people write documents, presentations and emails, Microsoft has discussed incorporating OpenAI's artificial intelligence in Word, PowerPoint, Outlook and other apps so customers can automatically generate text using simple prompts,according to ...

Short Title	Ghost Writer
URL	https://www.theinformation.com/articles/ghost-writer-microsoft-looks-to-add-openais-chatbot-technology-to-word-email
Accessed	1/7/2023, 11:19:39 AM
Extra	WFZ Note the discussions of training GPT on customer data. This makes all the engineering way more complicated. Also raises a lot of product and substantive questions. Do I actually want my writing to be more like myself? Do I want it to be more like what is found on my corporate intranet? I don't think these are foregone conclusions.
Publication	The Information
Date Added	1/7/2023, 11:19:39 AM
Modified	1/7/2023, 11:35:01 AM

- # Attachments

  - ## Snapshot

- # Google and Universal Music negotiate deal over AI 'deepfakes' | Financial Times

Item Type	Web Page
URL	https://www.ft.com/content/6f022306-2f83-4da7-8066-51386e8fe63b

Accessed	8/10/2023, 4:04:04 AM
Extra	WFZ First in what will be a long series of pairings off between traditional media companies and a handful of major AI companies. At the end of phase 1, everyone will have a partner. At the end of phase 2, the traditional media companies will be more like agents who bundle rights while the AI companies produce all the art.
Date Added	8/10/2023, 4:04:04 AM
Modified	8/10/2023, 5:35:37 AM

- # Attachments

  - Google and Universal Music negotiate deal over AI 'deepfakes' | Financial Times

- # Google Tests an A.I. Assistant That Offers Life Advice

Item Type	Newspaper Article
Author	Nico Grant
Abstract	The tech giant is evaluating tools that would use artificial intelligence to perform tasks that some of its researchers have said should be avoided.
Date	2023-08-16
Language	en-US

Library Catalog	NYTimes.com
URL	https://www.nytimes.com/2023/08/16/technology/google-ai-life-advice.html
Accessed	8/16/2023, 6:52:55 PM
Section	Technology
Publication	The New York Times
ISSN	0362-4331
Date Added	8/16/2023, 6:52:55 PM
Modified	8/16/2023, 6:53:00 PM

- # Tags:

  - Artificial Intelligence
  - Computers and the Internet
  - DeepMind Technologies Ltd
  - Bard
  - ChatGPT
  - Google DeepMind
  - Google Inc
  - Mergers, Acquisitions and Divestitures

Notes:

- Google DeepMind has also been evaluating tools recently that could take its A.I. further into the workplace, including capabilities to generate scientific, creative and professional writing, as well as to recognize patterns and extract data from text, according to the documents, potentially making it relevant to knowledge workers in various industries and fields. The company's A.I. safety experts had also expressed concern about the economic harms of generative A.I. in the December presentation reviewed by The Times, arguing that it could lead to the "deskilling of creative writers."

Attachments

- Snapshot

- # GPT-3 Powers the Next Generation of Apps

Item Type	Web Page
Author	OpenAI
Date	2021
URL	https://openai.com/blog/gpt-3-apps/

Accessed	3/28/2021, 8:00:00 PM
Date Added	3/29/2021, 6:48:39 PM
Modified	12/4/2022, 11:50:38 PM

- ## Tags:

  - #nosource

- # GPT-4

Item Type	Web Page
Abstract	We've created GPT-4, the latest milestone in OpenAI's effort in scaling up deep learning. GPT-4 is a large multimodal model (accepting image and text inputs, emitting text outputs) that, while less capable than humans in many real-world scenarios, exhibits human-level performance on various professional and academic benchmarks.
Language	en-US
URL	https://openai.com/research/gpt-4
Accessed	3/15/2023, 1:02:04 AM
Date Added	3/15/2023, 1:02:04 AM
Modified	3/15/2023, 1:02:04 AM

- **Attachments**

  - gpt-4.pdf
  - Snapshot

gpt-author

Item Type	Software
Programmer	mshumer
Date	2023-06-21T09:46:10Z
Library Catalog	GitHub
URL	https://github.com/mshumer/gpt-author
Accessed	6/21/2023, 5:51:59 AM
Rights	MIT
Extra	WFZ Simple fantasy novel generator open source, written by Matt Shumer of Otherside AI.
Prog. Language	Jupyter Notebook
Date Added	6/21/2023, 5:51:59 AM
Modified	6/21/2023, 10:55:11 PM

- # GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models

Item Type	Preprint
Author	Tyna Eloundou
Author	Sam Manning
Author	Pamela Mishkin
Author	Daniel Rock

Abstract — We investigate the potential implications of large language models (LLMs), such as Generative Pre-trained Transformers (GPTs), on the U.S. labor market, focusing on the increased capabilities arising from LLM-powered software compared to LLMs on their own. Using a new rubric, we assess occupations based on their alignment with LLM capabilities, integrating both human expertise and GPT-4 classifications. Our findings reveal that around 80% of the U.S. workforce could have at least 10% of their work tasks affected by the introduction of LLMs, while approximately 19% of workers may see at least 50% of their tasks impacted. We do not make predictions about the development or adoption timeline of such LLMs. The projected effects span all wage levels, with higher-income jobs potentially facing greater exposure to LLM capabilities and LLM-powered software. Significantly, these impacts are not restricted to industries with higher recent productivity growth. Our analysis suggests that, with access to an LLM, about 15% of all worker tasks in the US could be completed significantly faster at the same level of quality. When incorporating software and tooling built on top of LLMs, this share increases to between 47 and 56% of all tasks. This finding implies that LLM-powered software will have a substantial effect on scaling the economic impacts of the underlying models. We conclude that LLMs such as GPTs exhibit traits of general-purpose technologies, indicating that they could have considerable economic, social, and policy implications.

Date	2023-03-23
Short Title	GPTs are GPTs
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2303.10130
Accessed	4/23/2023, 2:44:41 AM
Extra	WFZ Book publishing had the fifth highest exposure to LLM task replacement out of 96 industries evaluated. See p. 28.
DOI	10.48550/arXiv.2303.10130
Repository	arXiv
Archive ID	arXiv:2303.10130
Date Added	4/23/2023, 2:44:42 AM
Modified	4/23/2023, 2:59:00 AM

- # Tags:

  - Computer Science - Artificial Intelligence
  - Computer Science - Computers and Society
  - Economics - General Economics

Attachments

Graph-based Semantical Extractive Text Analysis

Item Type	Preprint
Author	Mina Samizadeh
Abstract	In the past few decades, there has been an explosion in the amount of available data produced from various sources with different topics. The availability of this enormous data necessitates us to adopt effective computational tools to explore the data. This leads to an intense growing interest in the research community to develop computational methods focused on processing this text data. A line of study focused on condensing the text so that we are able to get a higher level of understanding in a shorter time. The two important tasks to do this are keyword extraction and text summarization. In keyword extraction, we are interested in finding the key important words from a text. This makes us familiar with the general topic of a text. In text summarization, we are interested in producing a short-length text which includes important information about the document. The TextRank algorithm, an unsupervised learning method that is an extension of the PageRank (algorithm which is the base algorithm of Google search engine for searching pages and ranking them) has shown its efficacy in large-scale text mining, especially for text summarization and keyword extraction. this algorithm can automatically extract the important parts of a text (keywords or sentences) and declare them as the result. However, this algorithm neglects the semantic similarity between the different parts. In this work, we improved the results of the TextRank algorithm by incorporating the semantic similarity between parts of the text. Aside from keyword extraction and text summarization, we develop a topic clustering algorithm based on our framework which can be used individually or as a part of generating the summary to overcome coverage problems.
Date	2022-12-19
Language	en

Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2212.09701
Accessed	12/21/2022, 11:49:10 AM
Extra	WFZ Explores combination of semantic and graph based summarization techniques. One yellow flag is that the reference implmentation at github relies on gensim's text summarization implementation, which IIRC was deprecated because the gensim maintainers did not have confidence it was adding value.
Repository	arXiv
Archive ID	arXiv:2212.09701
Date Added	12/21/2022, 11:49:11 AM
Modified	12/21/2022, 12:19:49 PM

- **Tags:**

    - Computer Science - Computation and Language
    - Computer Science - Machine Learning

Attachments

    - Samizadeh - 2022 - Graph-based Semantical Extractive Text Analysis.pdf

- Greg Brockman on Twitter

Item Type	Web Page
Author	Greg Brockman
Abstract	"Exciting but overlooked that ChatGPT is primarily an alignment advance—the base model (GPT-3.5) has been available in publicly for many months, but making it into a useful chat system required significant strides with reliably following the intent of the developer and the user."
Date	2022-12-04
Language	en
URL	https://twitter.com/gdb/status/1599124287633248257
Accessed	12/4/2022, 9:34:45 PM
Extra	WFZ OpenAI founder Greg Brockman confirms that chatGPT is mainly an improvement in alignment & productization. The LLM is the same as GPT3.5 has been for months.
Website Title	Twitter
Date Added	12/4/2022, 9:34:45 PM
Modified	12/5/2022, 12:24:34 AM

- ## **Notes:**

  - ○ Greg Brockman @gdb Exciting but overlooked that ChatGPT is primarily an alignment advance—the base model (GPT-3.5) has been

available in publicly for many months, but making it into a useful chat system required significant strides with reliably following the intent of the developer and the user.

Attachments

- ○ Snapshot

Guiding Large Language Models via Directional Stimulus Prompting

Item Type	Preprint
Author	Zekun Li
Author	Baolin Peng
Author	Pengcheng He
Author	Michel Galley
Author	Jianfeng Gao
Author	Xifeng Yan
Abstract	We introduce a new framework, Directional Stimulus Prompting, that uses a tuneable language model (LM) to provide guidance for the black-box frozen large language model (LLM) on downstream tasks. Unlike prior work that manually or automatically finds the optimal prompt for each task, we train a policy LM to generate discrete tokens as ``directional stimulus'' of each input, which is a hint/cue such as keywords of an article for summarization. The directional stimulus is then combined with the original input and fed into the LLM to guide its generation toward the desired target. The policy LM can be

trained through 1) supervised learning from annotated data and 2) reinforcement learning from offline and online rewards to explore directional stimulus that better aligns LLMs with human preferences. This framework is flexibly applicable to various LMs and tasks. To verify its effectiveness, we apply our framework to summarization and dialogue response generation tasks. Experimental results demonstrate that it can significantly improve LLMs' performance with a small collection of training data: a T5 (780M) trained with 2,000 samples from the CNN/Daily Mail dataset improves Codex (175B)'s performance by 7.2% in ROUGE-Avg scores; 500 dialogues boost the combined score by 52.5%, achieving comparable or even better performance than fully trained models on the MultiWOZ dataset.

Date	2023-02-22
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2302.11520
Accessed	2/25/2023, 10:15:35 AM
Extra	arXiv:2302.11520 [cs]
Repository	arXiv
Archive ID	arXiv:2302.11520
Date Added	2/25/2023, 10:15:35 AM
Modified	2/25/2023, 10:15:35 AM

- # Tags:

  - Computer Science - Computation and Language

Attachments

- o **arXiv Fulltext PDF**
- o **arXiv.org Snapshot**

- # Harvard psychologist: If you use any of these 9 phrases every day, 'you're more emotionally resilient than most'

Item Type	Web Page
Author	Dr Cortney Warren Contributor
Abstract	Very few of us are able to quickly adapt to stressful situations. Harvard psychologist Dr. Courtney Warren shares the powerful phrases that the most emotionally resilient people say every day.
Date	2023-08-06
Language	en
Short Title	Harvard psychologist
URL	https://www.cnbc.com/2023/08/06/harvard-psychologist-if-you-use-any-of-these-phrases-you-are-more-emotionally-resilient-than-most.html
Accessed	8/10/2023, 10:45:11 PM
Website Title	CNBC

Date Added	8/10/2023, 10:45:10 PM
Modified	8/10/2023, 10:45:10 PM

- ## **Attachments**

  - ○ Snapshot

- # High-Resolution Image Synthesis with Latent Diffusion Models

Item Type	Preprint
Author	Robin Rombach
Author	Andreas Blattmann
Author	Dominik Lorenz
Author	Patrick Esser
Author	Björn Ommer
Abstract	By decomposing the image formation process into a sequential application of denoising autoencoders, diffusion models (DMs) achieve state-of-the-art synthesis results on image data and beyond. Additionally, their formulation allows for a guiding mechanism to control the image generation process without retraining. However, since these models typically operate directly in pixel space, optimization of powerful DMs often consumes hundreds of GPU days and inference is expensive due to sequential evaluations. To enable DM training on limited computational resources while retaining their quality and flexibility, we apply them in the latent space of powerful pretrained autoencoders. In contrast to previous work, training diffusion models on such a representation

allows for the first time to reach a near-optimal point between complexity reduction and detail preservation, greatly boosting visual fidelity. By introducing cross-attention layers into the model architecture, we turn diffusion models into powerful and flexible generators for general conditioning inputs such as text or bounding boxes and high-resolution synthesis becomes possible in a convolutional manner. Our latent diffusion models (LDMs) achieve a new state of the art for image inpainting and highly competitive performance on various tasks, including unconditional image generation, semantic scene synthesis, and super-resolution, while significantly reducing computational requirements compared to pixel-based DMs. Code is available at https://github.com/CompVis/latent-diffusion .

Date	2022-04-13
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2112.10752
Accessed	10/2/2022, 12:00:18 PM
Extra	WFZ Documents Stable Diffusion model.
DOI	10.48550/arXiv.2112.10752
Repository	arXiv
Archive ID	arXiv:2112.10752
Date Added	10/2/2022, 12:00:18 PM
Modified	10/2/2022, 12:01:58 PM

- # **Tags:**

- o Computer Science - Computer Vision and Pattern Recognition

Notes:

- o Comment: CVPR 2022

Attachments

- o arXiv Fulltext PDF
- o arXiv.org Snapshot

- # Holistic Evaluation of Language Models

Item Type	Preprint
Author	Percy Liang
Author	Rishi Bommasani
Author	Tony Lee
Author	Dimitris Tsipras
Author	Dilara Soylu
Author	Michihiro Yasunaga

Author   Yian Zhang

Author   Deepak Narayanan

Author   Yuhuai Wu

Author   Ananya Kumar

Author   Benjamin Newman

Author   Binhang Yuan

Author   Bobby Yan

Author   Ce Zhang

Author   Christian Cosgrove

Author   Christopher D. Manning

Author   Christopher Ré

Author   Diana Acosta-Navas

Author   Drew A. Hudson

Author   Eric Zelikman

Author   Esin Durmus

Author   Faisal Ladhak

Author   Frieda Rong

Author   Hongyu Ren

Author   Huaxiu Yao

Author   Jue Wang

Author   Keshav Santhanam

Author   Laurel Orr

Author   Lucia Zheng

Author   Mert Yuksekgonul

Author   Mirac Suzgun

Author   Nathan Kim

Author   Neel Guha

Author   Niladri Chatterji

Author   Omar Khattab

Author   Peter Henderson

Author   Qian Huang

Author   Ryan Chi

Author   Sang Michael Xie

Author   Shibani Santurkar

Author   Surya Ganguli

Author   Tatsunori Hashimoto

Author   Thomas Icard

Author   Tianyi Zhang

Author   Vishrav Chaudhary

Author   William Wang

Author   Xuechen Li

Author   Yifan Mai

Author   Yuhui Zhang

Author   Yuta Koreeda

Abstract   Language models (LMs) are becoming the foundation for almost all major language technologies, but their capabilities, limitations, and risks are not well understood. We present Holistic Evaluation of Language Models (HELM) to improve the transparency of language models. First, we taxonomize the vast space of potential scenarios (i.e. use cases) and metrics (i.e. desiderata) that are of interest for LMs. Then we select a broad subset based on coverage and feasibility, noting what's missing or underrepresented (e.g. question answering for neglected English dialects, metrics for trustworthiness). Second, we adopt a multi-metric approach: We measure 7 metrics (accuracy, calibration, robustness, fairness, bias, toxicity, and efficiency) for each of 16 core scenarios when possible (87.5% of the time). This ensures metrics beyond accuracy don't fall to the wayside, and that trade-offs are clearly exposed. We also perform 7 targeted evaluations, based on 26 targeted scenarios, to analyze specific aspects (e.g. reasoning, disinformation). Third, we conduct a large-scale evaluation of 30 prominent language models (spanning open, limited-access, and closed models) on all 42 scenarios, 21 of which were not previously used in mainstream LM evaluation. Prior to HELM, models on average were evaluated on just 17.9% of the core HELM scenarios, with some prominent models not sharing a single

scenario in common. We improve this to 96.0%: now all 30 models have been densely benchmarked on the same core scenarios and metrics under standardized conditions. Our evaluation surfaces 25 top-level findings. For full transparency, we release all raw model prompts and completions publicly for further analysis, as well as a general modular toolkit. We intend for HELM to be a living benchmark for the community, continuously updated with new scenarios, metrics, and models.

Date	2022-11-16
Language	en
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2211.09110
Accessed	11/24/2022, 7:56:48 AM
Extra	WFZ Major paper from the Center for Research on Foundation Models (CRFM) at Stanford attempts to provide a comprehensive framework for evaluation of *all* language models. Long paper (163 p.) Money shots (head-to-head results) begin around p.51 with the key graph at page 55. OpenAI's GPT-3 davinci instruct comes out near the top on every metric. Cohere is a somewhat surprising #2. Will repay close reading with many interesting details on the relationships among metrics and prompts.
Repository	arXiv
Archive ID	arXiv:2211.09110
Date Added	11/24/2022, 7:56:48 AM
Modified	11/24/2022, 12:55:08 PM

- # **Tags:**

- ○ Computer Science - Computation and Language
- ○ Computer Science - Machine Learning
- ○ Computer Science - Artificial Intelligence

Notes:

- ○ Comment: Authored by the Center for Research on Foundation Models (CRFM) at the Stanford Institute for Human-Centered Artificial Intelligence (HAI). Project page: https://crfm.stanford.edu/helm/v1.0

Attachments

- ○ Liang et al. - 2022 - Holistic Evaluation of Language Models.pdf

- # Hopes for 2023 from Yoshua Bengio, Been Kim, Douwe Kiela, Reza Zadeh, Alon Halevy

Item Type	Web Page
Abstract	As we enter the new year, let's view 2023 not as a single year, but as the first of more in which we will accomplish our long-term goals. Some results take a long time to achieve...
Date	2022-12-28T12:29:13.000-08:00

Language	en
URL	https://www.deeplearning.ai/the-batch/issue-177/
Accessed	1/3/2023, 9:14:05 PM
Extra	WFZ Reza Zadeh: "with recent advances in generative AI for images and text, active learning is primed for a major breakthrough. Now, when a learning algorithm is unsure of the correct label for some part of its encoding space, it can actively generate data from that section to get input from a human." AIs giving homework assignments. It must be said that the scenarios for this in book publishing are pretty dystopian. AIs generating slushpiles for editorial assistants to review. Yikes!
Website Title	Hopes for 2023 from Yoshua Bengio, Been Kim, Douwe Kiela, Reza Zadeh, Alon Halevy
Date Added	1/3/2023, 9:14:05 PM
Modified	1/3/2023, 9:22:48 PM

- # Notes:

  - Now, when a learning algorithm is unsure of the correct label for some part of its encoding space, it can actively generate data from that section to get input from a human.

Attachments

- Snapshot

- # How A Book About Flies Came To Be Priced $24 Million On Amazon | WIRED

Item Type	Web Page
URL	https://www.wired.com/2011/04/amazon-flies-24-million/
Accessed	8/27/2022, 2:05:13 AM
Date Added	8/27/2022, 2:05:13 AM
Modified	8/27/2022, 2:05:13 AM

- # Attachments

  - How A Book About Flies Came To Be Priced $24 Million On Amazon | WIRED

- # How Algorithmically Created Content will Transform Publishing

Item Type	Web Page
Author	Dan Woods
Abstract	Fred Zimmerman, who is CEO of Nimble Books, is pioneering a new technique he calls combinatorial publishing that can create a book that is useful in seconds for pennies. He persuasively argues that algorithmic content creation has an important

role to play, even if the virtuosity of the human will always be the beating heart of content creation.

Language	en
URL	https://www.forbes.com/sites/danwoods/2012/08/13/how-algorithmically-created-content-will-transform-publishing/
Accessed	3/15/2023, 2:50:57 PM
Extra	WFZ An overnight success, ten years in.
Website Title	Forbes
Date Added	3/15/2023, 2:50:57 PM
Modified	3/15/2023, 2:51:54 PM

- ## Attachments

  - Snapshot

- # How an AI-written Star Wars story created chaos at Gizmodo

Item Type	Newspaper Article
Author	Pranshu Verma

Abstract   A Gizmodo story on Star Wars, generated by artificial intelligence, was riddled with errors. The irony that the problem happened at a tech publication was undeniable.

Date   2023-07-08

Language   en-US

Library Catalog   www.washingtonpost.com

URL   https://www.washingtonpost.com/technology/2023/07/08/gizmodo-ai-errors-star-wars/

Accessed   7/8/2023, 5:48:22 PM

Publication   Washington Post

ISSN   0190-8286

Date Added   7/8/2023, 5:48:22 PM

Modified   7/8/2023, 5:48:22 PM

- # Notes:

  - The media watchdog NewsGuard said that at least 301 AI-generated news sites exist that operate with "no human oversight and publish articles written largely or entirely by bots," and span 13 languages, including English, Arabic, Chinese and French.

Attachments

- Snapshot

How Far Can Camels Go? Exploring the State of Instruction Tuning on Open Resources

Item Type	Preprint
Author	Yizhong Wang
Author	Hamish Ivison
Author	Pradeep Dasigi
Author	Jack Hessel
Author	Tushar Khot
Author	Khyathi Raghavi Chandu
Author	David Wadden
Author	Kelsey MacMillan
Author	Noah A. Smith
Author	Iz Beltagy
Author	Hannaneh Hajishirzi

Abstract   In this work we explore recent advances in instruction-tuning language models on a range of open instruction-following datasets. Despite recent claims that open models can be on par with state-of-the-art proprietary models, these claims are often accompanied by limited evaluation, making it difficult to compare models across the board and determine the utility of various resources. We provide a large set of instruction-tuned models from 6.7B to 65B parameters in size, trained on 12 instruction datasets ranging from manually curated (e.g., OpenAssistant) to synthetic and distilled (e.g., Alpaca) and systematically evaluate them on their factual knowledge, reasoning, multilinguality, coding, and open-ended instruction following abilities through a collection of automatic, model-based, and human-based metrics. We further introduce T\"ulu, our best performing instruction-tuned model suite finetuned on a combination of high-quality open resources. Our experiments show that different instruction-tuning datasets can uncover or enhance specific skills, while no single dataset (or combination) provides the best performance across all evaluations. Interestingly, we find that model and human preference-based evaluations fail to reflect differences in model capabilities exposed by benchmark-based evaluations, suggesting the need for the type of systemic evaluation performed in this work. Our evaluations show that the best model in any given evaluation reaches on average 83% of ChatGPT performance, and 68% of GPT-4 performance, suggesting that further investment in building better base models and instruction-tuning data is required to close the gap. We release our instruction-tuned models, including a fully finetuned 65B T\"ulu, along with our code, data, and evaluation framework at https://github.com/allenai/open-instruct to facilitate future research.

Date   2023-06-07

Short Title   How Far Can Camels Go?

Library Catalog   arXiv.org

URL   http://arxiv.org/abs/2306.04751

Accessed   7/22/2023, 6:33:50 PM

Extra   arXiv:2306.04751 [cs]

DOI   10.48550/arXiv.2306.04751

- ## Tags:

  - Computer Science - Computation and Language

Notes:

  - Comment: 18 pages, 5 figure, 7 tables. Under the review of NeurIPS 2023 Datasets and Benchmarks Track

Attachments

  - arXiv Fulltext PDF
  - arXiv.org Snapshot

- # How Generative AI Can Augment Human Creativity

Item Type	Magazine Article

Author	Tojin T. Eapen
Author	Daniel J. Finkenstadt
Author	Josh Folk
Author	Lokesh Venkataswamy
Abstract	There is tremendous apprehension about the potential of generative AI—technologies that can create new content such as text, images, and video—to replace people in many jobs. But one of the biggest opportunities generative AI offers is to augment human creativity and overcome the challenges of democratizing innovation. In the past two decades, companies have used crowdsourcing and idea competitions to involve outsiders in the innovation process. But many businesses have struggled to capitalize on these contributions. They've lacked an efficient way to evaluate the ideas, for instance, or to synthesize different ideas. Generative AI can help overcome those challenges, the authors say. It can supplement the creativity of employees and customers and help them produce and identify novel ideas—and improve the quality of raw ideas. Specifically, companies can use generative AI to promote divergent thinking, challenge expertise bias, assist in idea evaluation, support idea refinement, and facilitate collaboration among users.
Date	2023-07-01T04:00:00Z
Library Catalog	hbr.org
URL	https://hbr.org/2023/07/how-generative-ai-can-augment-human-creativity
Accessed	8/9/2023, 4:30:55 AM
Extra	Section: AI and machine learning
Publication	Harvard Business Review
ISSN	0017-8012

Date Added	8/9/2023, 4:30:55 AM
Modified	8/9/2023, 4:30:55 AM

- ## Tags:

  - AI and machine learning
  - Creativity
  - Innovation

Attachments

  - Snapshot

- # How has AI developed over the years and what's next?

Item Type	Document
Author	World Economic Forum
Abstract	Artificial intelligence has come a long way since the 1950s. We now have AI systems like DALL-E and PaLM with abilities to produce photorealistic images and interpret and generate language.
Language	en
URL	https://www.weforum.org/agenda/2022/12/how-ai-developed-whats-next-digital-transformation/

Accessed	12/23/2022, 5:36:00 PM
Extra	WFZ Good overview. Includes widely reproduced image of FLOPs growth: https://ourworldindata.org/uploads/2022/12/ai-training-computation-3.png.
Date Added	12/23/2022, 5:36:00 PM
Modified	12/26/2022, 12:27:43 AM

- ## **Attachments**

  - ○ Snapshot

- # How Nvidia dominated AI — and plans to keep it that way as generative AI explodes

Item Type	Blog Post
Abstract	Over the past decade, Nvidia has dominated the AI landscape. Can it handle 2023's challenges and win big in the generative AI gold rush?
Date	2023-02-23T13:00:00+00:00
Language	en-US
URL	https://venturebeat.com/ai/how-nvidia-dominated-ai-and-plans-to-keep-it-that-way-as-generative-ai-explodes/
Accessed	2/27/2023, 10:42:30 PM

Extra	WFZ Good overview that provides valuable context about a key company that is probably not super well known to most people in book publishing.
Blog Title	VentureBeat
Date Added	2/27/2023, 10:42:30 PM
Modified	2/27/2023, 10:43:29 PM

- ## Attachments

  - ### Snapshot

- # How TikTok Became a Best-Seller Machine

Item Type	Web Page
Abstract	#BookTok, where enthusiastic readers share reading recommendations, has gone from being a novelty to becoming an anchor in the publishing industry and a dominant driver of fiction sales.
Date	2022-07-01T13:38:09.000Z
Language	en
URL	https://www.nytimes.com/2022/07/01/books/tiktok-books-booktok.html
Accessed	9/26/2022, 10:51:49 PM
Extra	Section: Books

Date Added	9/26/2022, 10:51:49 PM
Modified	9/26/2022, 10:51:53 PM

- ## **Attachments**

  - Snapshot

- # How to evaluate OpenAI's super dialogue model ChatGPT?

Item Type	Forum Post
Author	Tian Yuandong
Date	2022-12-04
URL	如何评价OpenAI的超级对话模型ChatGPT？- 周博磊的回答 - 知乎 https://www.zhihu.com/question/570189639/answer/2804667785
Extra	WFZ Facebook researcher experimenting with ChatGPT as a tool for writing novels. Key finding: some deep imagination and connection ability are still missing.If you want to write a more interesting plot, it seems that the key part of the plot still needs the author to develop and induce the AI system to complete [machine translation]
Forum/Listserv Title	Zhihu
Date Added	12/19/2022, 2:53:52 PM

How to Find Comp Titles Using ChatGPT

Item Type	Blog Post
Author	John Matthew Fox
Abstract	These five steps will help you find your ideal comp titles for your query letter or book proposal, using ChatGPT. Includes sample prompts.
Date	2023-04-07T09:00:00+00:00
Language	en-US
URL	https://www.janefriedman.com/how-to-find-comp-titles-using-chatgpt/
Accessed	4/7/2023, 9:30:30 PM
Extra	WFZ A nice technique. Would be even better if linked to a Nielsen BookScan ChatGPT plugin. May be worth tweaking prompts to see if it can arrive at real numbers.
Blog Title	Jane Friedman
Date Added	4/7/2023, 9:30:30 PM
Modified	4/8/2023, 6:33:55 PM

- ## Notes:

- Sample prompts for ChatGPT Suggest 10 comp titles published in the last three years for a literary novel set in New York that features drug use. Do not include any books that have appeared on bestseller lists. Do not include nonfiction books.

Attachments

- Snapshot

- # How to Get Access to the Midjourney API — Tokenized

Item Type	Blog Post
Author	Christian Heidorn
URL	https://tokenizedhq.com/midjourney-api/
Accessed	1/24/2023, 11:08:22 PM
Extra	WFZ tl;dr: As of Jan 2023, there is no public API. Good investigation.
Blog Title	Tokenized
Date Added	1/24/2023, 11:08:22 PM
Modified	1/24/2023, 11:10:04 PM

- **Attachments**

  o How to Get Access to the Midjourney API — Tokenized

How to Read a Paper: The Basics of Evidence-based Medicine and Healthcare

Item Type	Book
Author	Trisha Greenhalgh
Abstract	Required reading in many medical and healthcare institutions, How to Read a Paper is a clear and wide-ranging introduction to evidence-based medicine and healthcare, helping readers to understand its central principles, critically evaluate published data, and implement the results in practical settings. Author Trisha Greenhalgh guides readers through each fundamental step of inquiry, from searching the literature to assessing methodological quality and appraising statistics.How to Read a Paper addresses the common criticisms of evidence-based healthcare, dispelling many of its myths and misconceptions, while providing a pragmatic framework for testing the validity of healthcare literature. Now in its sixth edition, this informative text includes new and expanded discussions of study bias, political interference in published reports, medical statistics, big data and more.Offers user-friendly guidance on evidence-based healthcare that is applicable to both experienced and novice readersAuthored by an internationally recognised practitioner and researcher in evidence-based healthcare and primary careIncludes updated references, additional figures, improved checklists and moreHow to Read a Paper is an ideal resource for healthcare students, practitioners and anyone seeking an accessible introduction to evidence-based healthcare.
Date	2019-05-06
Language	English

Short Title	How to Read a Paper
Library Catalog	Amazon
Extra	WFZ A worthy modern successor to gems like Mortimer Adler's HOW TO READ A BOOK.
Place	Hoboken, NJ
Publisher	Wiley-Blackwell
ISBN	978-1-119-48474-5
Edition	6th edition
# of Pages	288
Date Added	3/12/2023, 2:05:59 PM
Modified	3/15/2023, 4:54:52 AM

- # Attachments

  - Amazon.com Link

- # https://twitter.com/mathemagic1an/status/1617620133182332928

Item Type	Web Page

Language	en
Short Title	https
URL	https://twitter.com/mathemagic1an/status/1617620133182332928
Accessed	1/23/2023, 6:15:45 PM
Website Title	Twitter
Date Added	1/23/2023, 6:15:53 PM
Modified	1/23/2023, 6:15:53 PM

• Huberman AI

Item Type	Web Page
Abstract	Use AI to explore the wisdom of The Huberman Lab.
Language	en
URL	https://huberman.rile.yt/
Accessed	12/28/2022, 4:38:58 AM
Extra	WFZ Every book and every author should have question-answering AIs like this one.
Website Title	Huberman AI

Date Added	12/28/2022, 4:38:58 AM
Modified	12/28/2022, 6:06:17 PM

- ## **Attachments**

  - ○ Snapshot

- # HuggingFace's Transformers: State-of-the-art Natural Language Processing

Item Type	Preprint
Author	Thomas Wolf
Author	Lysandre Debut
Author	Victor Sanh
Author	Julien Chaumond
Author	Clement Delangue
Author	Anthony Moi
Author	Pierric Cistac
Author	Tim Rault

Author   Rémi Louf

Author   Morgan Funtowicz

Author   Joe Davison

Author   Sam Shleifer

Author   Patrick von Platen

Author   Clara Ma

Author   Yacine Jernite

Author   Julien Plu

Author   Canwen Xu

Author   Teven Le Scao

Author   Sylvain Gugger

Author   Mariama Drame

Author   Quentin Lhoest

Author   Alexander M. Rush

Abstract   Recent progress in natural language processing has been driven by advances in both model architecture and model pretraining. Transformer architectures have facilitated building higher-capacity models and pretraining has made it possible to effectively utilize this capacity for a wide variety of tasks. \textit{Transformers} is an open-source library with the goal of opening up these advances to the wider machine learning community. The library consists of carefully engineered state-of-the art Transformer architectures under a

unified API. Backing this library is a curated collection of pretrained models made by and available for the community. \textit{Transformers} is designed to be extensible by researchers, simple for practitioners, and fast and robust in industrial deployments. The library is available at \url{https://github.com/huggingface/transformers}.

Date	2020-07-13
Short Title	HuggingFace's Transformers
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/1910.03771
Accessed	9/11/2022, 3:13:29 PM
Extra	arXiv:1910.03771 [cs]
DOI	10.48550/arXiv.1910.03771
Repository	arXiv
Archive ID	arXiv:1910.03771
Date Added	9/11/2022, 3:13:29 PM
Modified	9/11/2022, 3:13:32 PM

- ## Tags:

    - Computer Science - Computation and Language

Notes:

- ○ Comment: 8 pages, 4 figures, more details at https://github.com/huggingface/transformers

Attachments

- ○ arXiv Fulltext PDF
- ○ arXiv.org Snapshot

HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in HuggingFace

Item Type	Preprint
Author	Yongliang Shen
Author	Kaitao Song
Author	Xu Tan
Author	Dongsheng Li
Author	Weiming Lu
Author	Yueting Zhuang
Abstract	Solving complicated AI tasks with different domains and modalities is a key step toward artificial general intelligence (AGI). While there are abundant AI models available for different domains and modalities, they cannot handle

complicated AI tasks. Considering large language models (LLMs) have exhibited exceptional ability in language understanding, generation, interaction, and reasoning, we advocate that LLMs could act as a controller to manage existing AI models to solve complicated AI tasks and language could be a generic interface to empower this. Based on this philosophy, we present HuggingGPT, a system that leverages LLMs (e.g., ChatGPT) to connect various AI models in machine learning communities (e.g., HuggingFace) to solve AI tasks. Specifically, we use ChatGPT to conduct task planning when receiving a user request, select models according to their function descriptions available in HuggingFace, execute each subtask with the selected AI model, and summarize the response according to the execution results. By leveraging the strong language capability of ChatGPT and abundant AI models in HuggingFace, HuggingGPT is able to cover numerous sophisticated AI tasks in different modalities and domains and achieve impressive results in language, vision, speech, and other challenging tasks, which paves a new way towards AGI.

Date	2023-03-30
Short Title	HuggingGPT
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2303.17580
Accessed	3/31/2023, 12:40:18 PM
Extra	WFZ An example of using AI as an agent to handle complex decision making. Imagine a similar flow for publishers.
DOI	10.48550/arXiv.2303.17580
Repository	arXiv
Archive ID	arXiv:2303.17580
Date Added	3/31/2023, 12:40:18 PM

<table>
<tr><td>Modified</td><td>3/31/2023, 1:03:48 PM</td></tr>
</table>

- **Tags:**

  - Computer Science - Computation and Language
  - Computer Science - Machine Learning
  - Computer Science - Artificial Intelligence
  - Computer Science - Computer Vision and Pattern Recognition

Attachments

  - arXiv Fulltext PDF
  - arXiv.org Snapshot

- # Hungry Hungry Hippos: Towards Language Modeling with State Space Models

<table>
<tr><td>Item Type</td><td>Preprint</td></tr>
<tr><td>Author</td><td>Tri Dao</td></tr>
<tr><td>Author</td><td>Daniel Y. Fu</td></tr>
<tr><td>Author</td><td>Khaled K. Saab</td></tr>
<tr><td>Author</td><td>Armin W. Thomas</td></tr>
</table>

Author	Atri Rudra
Author	Christopher Ré

Abstract State space models (SSMs) have demonstrated state-of-the-art sequence modeling performance in some modalities, but underperform attention in language modeling. Moreover, despite scaling nearly linearly in sequence length instead of quadratically, SSMs are still slower than Transformers due to poor hardware utilization. In this paper, we make progress on understanding the expressivity gap between SSMs and attention in language modeling, and on reducing the hardware barrier between SSMs and attention. First, we use synthetic language modeling tasks to understand the gap between SSMs and attention. We find that existing SSMs struggle with two capabilities: recalling earlier tokens in the sequence and comparing tokens across the sequence. To understand the impact on language modeling, we propose a new SSM layer, H3, that is explicitly designed for these abilities. H3 matches attention on the synthetic languages and comes within 0.4 PPL of Transformers on OpenWebText. Furthermore, a hybrid 125M-parameter H3-attention model that retains two attention layers surprisingly outperforms Transformers on OpenWebText by 1.0 PPL. Next, to improve the efficiency of training SSMs on modern hardware, we propose FlashConv. FlashConv uses a fused block FFT algorithm to improve efficiency on sequences up to 8K, and introduces a novel state passing algorithm that exploits the recurrent properties of SSMs to scale to longer sequences. FlashConv yields 2$\times$ speedup on the long-range arena benchmark and allows hybrid language models to generate text 1.6$\times$ faster than Transformers. Using FlashConv, we scale hybrid H3-attention language models up to 1.3B parameters on the Pile and find promising initial results, achieving lower perplexity than Transformers and outperforming Transformers in zero- and few-shot learning on a majority of tasks in the SuperGLUE benchmark.

Date	2022-12-28
Short Title	Hungry Hungry Hippos
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2212.14052

- **Tags:**

  - Computer Science - Computation and Language
  - Computer Science - Machine Learning

- i trained an ai chatbot on my childhood journal entries - so that i could engage in real-time dialogue with my "inner child" some reflections below:

Item Type	Forum Post
Author	michelle huang [@michellehuang42]

Date	2022-11-27T23:12Z
Language	en
URL	https://twitter.com/michellehuang42/status/1597005489413713921
Accessed	11/29/2022, 12:55:01 AM
Extra	WFZ Twitter thread describing a very clever and readily generalizable application of fine-tuning OpenAI GPT3. There's an old saying that there's a book in everyone -- now there's a book in every child.
Forum/Listserv Title	Twitter
Post Type	Tweet
Date Added	11/29/2022, 12:55:01 AM
Modified	11/29/2022, 12:59:35 AM

- # Attachments

  - ○ Snapshot

- # I Would Rather See My Books Get Pirated Than This (Or: Why Goodreads and Amazon Are Becoming Dumpster Fires)

Item Type	Blog Post
Author	Jane Friedman
Abstract	Amazon and Goodreads must take steps to combat the flood of AI-generated content that will mislead readers and damage author reputations.
Date	2023-08-07T09:00:00+00:00
Language	en-US
Short Title	I Would Rather See My Books Get Pirated Than This (Or
URL	https://janefriedman.com/i-would-rather-see-my-books-pirated/
Accessed	8/7/2023, 10:51:36 AM
Extra	WFZ "... Set On Fire". Prominent industry figure discovers especially deplorable scam leveraging her identity.
Blog Title	Jane Friedman
Date Added	8/7/2023, 10:51:36 AM
Modified	8/9/2023, 6:54:39 PM

- # Attachments

  - ○ Snapshot

- # I Wrote a Book with GPT-3 AI in 24 Hours — And Got It Published

Item Type	Blog Post
Author	Jukka Aalho
Abstract	In early 2021 I signed up for the GPT-3 beta program to see how good it is. A few days later I had co-authored Aum Golly—a book of AI…
Date	2023-02-03T17:21:10.315Z
Language	en
URL	https://medium.com/swlh/i-wrote-a-book-with-gpt-3-ai-in-24-hours-and-got-it-published-93cf3c96f120
Accessed	6/5/2023, 4:19:49 AM
Blog Title	The Startup
Date Added	6/5/2023, 4:19:49 AM
Modified	6/5/2023, 4:19:49 AM

- # I Wrote and Illustrated a Book in 12 Hours with ChatGPT and Midjourney–Again

Item Type	Blog Post
Author	Jukka Aalho
Abstract	In late 2022 I signed up for ChatGPT to see how good it is. A few days later I had co-authored Aum Golly 2 — a book of illustrated AI poems.
Date	2023-02-03T17:11:07.270Z
Language	en
URL	https://jukkaaalho.medium.com/i-wrote-and-illustrated-a-book-in-12-hours-with-chatgpt-and-midjourney-f5baa68ff646
Accessed	6/5/2023, 4:20:07 AM
Blog Title	Medium
Date Added	6/5/2023, 4:20:07 AM
Modified	6/5/2023, 4:20:07 AM

- # **Attachments**

  - ○ Snapshot

- # ICML 2023

Item Type	Web Page
URL	https://icml.cc/Conferences/2023/llm-policy

Accessed	1/5/2023, 9:26:40 PM
Extra	WFZ Detailed explanation of why the International Conference on Machine Learning decided to ban papers written entirely by AI for 2023. This is the source document, read this instead of relying upon news summaries. Nuanced and sensible. tldr: we don't want to be used for a publicity stunt.
Date Added	1/5/2023, 9:26:40 PM
Modified	1/7/2023, 5:09:43 PM

- # Attachments

  - ○ ICML 2023

- # Identify the Relevant Pages of Book to be Indexed Using Naive Bayes Classification Method

Item Type	Journal Article
Author	S. Christina
Author	D. Ronaldo
Abstract	This study aims to identify the relevance of the pages to be indexed in the back-of-book index, using Naive Bayes Classification, and shows that the naive bayes classification approach is capable of identifying the relevant and irrelevant pages. The back-of-book index is a component that is often found in non-fiction books. The back-of-book index contains important terms and page numbers where the terms appear in the book, it is useful for helping readers find certain term directly, without having to search on every page. Usually, the back-of-

book index is compiled by the author or a professional indexer, although currently there are several automatic indexing applications available. To determine a term on a page worth to be indexed requires knowledge and expertise of the author whom better understand the book's context. Hence, generating a good back-of-book index is a task that requires great effort, knowledge and cost. Therefore, the back-of-book index is very possible containing indexed terms that refer to irrelevant page numbers due to human error or the weakness of the indexing application. This study aims to identify the relevance of the pages to be indexed in the back-of-book index, using Naive Bayes Classification. The testing result shows that the approach with Naive Bayes Classification produces an average precision value of 74.02 percent and 100 percent for the recall value. The average precision value that more than 50 percent indicates that the naive bayes classification approach is capable to identify the relevant and irrelevant pages.

Date	2020
Library Catalog	Semantic Scholar
Publication	IOP Conference Series: Materials Science and Engineering
DOI	[10.1088/1757-899X/722/1/012043](10.1088/1757-899X/722/1/012043)
Date Added	8/24/2022, 3:19:29 PM
Modified	8/24/2022, 3:19:34 PM

- ## Attachments

  - Full Text
  - Semantic Scholar Link

- # Illustrating Reinforcement Learning from Human Feedback (RLHF)

Item Type	Web Page
Abstract	We're on a journey to advance and democratize artificial intelligence through open source and open science.
Date	2022-12-09
URL	https://huggingface.co/blog/rlhf
Accessed	12/9/2022, 9:32:49 PM
Extra	WFZ Nice package on RLHF with ELI5, graphic, and more.
Date Added	12/9/2022, 9:32:49 PM
Modified	12/22/2022, 3:05:05 AM

- ## Attachments

  - Snapshot

- # Image-making AIs are headed to your favorite fan universe

Item Type	Web Page
Author	Ina Fried
Abstract	All content owners should be able to set-up a generative AI system based on their intellectual property as a way to engage with fans, says Vermillio CEO Dan Neely. "It should happen, it should be easy and you should get paid," Neely said.

Date	2022-12-13
Language	en
Short Title	Axios Login
URL	https://www.axios.com/newsletters/axios-login/
Accessed	12/13/2022, 9:34:32 AM
Extra	WFZ This is right on target. The next step beyond is to make creating fan universes *in their entirety* -- both content & ancillary content -- a push-button enterprise. "MCU in a click." I'm working on this.
Website Title	Axios
Date Added	12/13/2022, 9:34:32 AM
Modified	12/15/2022, 8:44:56 AM

- **Tags:**

  - Character creation
  - Illustration

Notes:

  - Among those aiming to make that happen is Vermillio, a Chicago-based startup. The company has been in talks with a variety of

content owners and has signed at least one major deal, expected to be announced early next year. All content owners should be able to set-up a generative AI system based on their intellectual property as a way to engage with fans, says Vermillio CEO Dan Neely. "It should happen, it should be easy and you should get paid," Neely said.

Attachments

- ○ Snapshot

• Imagining better interfaces to language models | thesephist.com

Item Type	Web Page
URL	https://thesephist.com/posts/latent/
Accessed	10/13/2022, 4:37:41 PM
Extra	WFZ Insightful essay by Linus Lee discussing how user interfaces might evolve to help us explore the "latent spaces" of large language models.
Date Added	10/13/2022, 4:37:41 PM
Modified	10/13/2022, 11:04:07 PM

- ## Attachments

- o Imagining better interfaces to language models | thesephist.com

- # Impromptu :: Amplifying Our Humanity Through AI

Item Type	Web Page
Abstract	Impromptu: Amplifying our Humanity through AI, written by Reid Hoffman with GPT-4, takes readers on a travelog of the future – exploring how AI, and especially Large Language Models like GPT-4, can elevate humanity across key areas like education, business, and creativity.
Language	en-US
Short Title	Impromptu
URL	https://www.impromptubook.com/
Accessed	3/17/2023, 4:24:43 PM
Website Title	Impromptu
Date Added	3/17/2023, 4:24:43 PM
Modified	3/17/2023, 4:24:43 PM

- # Attachments

  - o Snapshot

- # Improving Automatic Quotation Attribution in Literary Novels

Item Type	Preprint
Author	Krishnapriya Vishnubhotla
Author	Frank Rudzicz
Author	Graeme Hirst
Author	Adam Hammond
Abstract	Current models for quotation attribution in literary novels assume varying levels of available information in their training and test data, which poses a challenge for in-the-wild inference. Here, we approach quotation attribution as a set of four interconnected sub-tasks: character identification, coreference resolution, quotation identification, and speaker attribution. We benchmark state-of-the-art models on each of these sub-tasks independently, using a large dataset of annotated coreferences and quotations in literary novels (the Project Dialogism Novel Corpus). We also train and evaluate models for the speaker attribution task in particular, showing that a simple sequential prediction model achieves accuracy scores on par with state-of-the-art models.
Date	2023-07-07
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2307.03734
Accessed	7/11/2023, 2:18:23 AM
Extra	arXiv:2307.03734 [cs]

- ## Tags:

  - Computer Science - Computation and Language

Notes:

  - Comment: Accepted to ACL 2023, short paper

Attachments

  - arXiv Fulltext PDF
  - arXiv.org Snapshot

- # Index, A History of the

Item Type	Book
Author	Dennis Duncan

Abstract	'Hilarious' Sam Leith'I loved this book' Susie Dent''Witty and affectionate' Lynne TrussPerfect for book lovers, a delightful history of the wonders to be found in the humble book indexMost of us give little thought to the back of the book - it's just where you go to look things up. But here, hiding in plain sight, is an unlikely realm of ambition and obsession, sparring and politicking, pleasure and play. Here we might find Butchers, to be avoided, or Cows that sh-te Fire, or even catch Calvin in his chamber with a Nonne. This is the secret world of the index: an unsung but extraordinary everyday tool, with an illustrious but little-known past. Here, for the first time, its story is told. Charting its curious path from the monasteries and universities of thirteenth-century Europe to Silicon Valley in the twenty-first, Dennis Duncan reveals how the index has saved heretics from the stake, kept politicians from high office and made us all into the readers we are today. We follow it through German print shops and Enlightenment coffee houses, novelists' living rooms and university laboratories, encountering emperors and popes, philosophers and prime ministers, poets, librarians and - of course - indexers along the way. Revealing its vast role in our evolving literary and intellectual culture, Duncan shows that, for all our anxieties about the Age of Search, we are all index-rakers at heart, and we have been for eight hundred years.*A Financial Times and History Today Book of the Year*
Date	2021-09-02
Language	en
Library Catalog	Google Books
Extra	WFZ Everything you need to know about the history of the index. In a nutshell, indexes began as a much larger genus, with dozens of floriant charismatic species, then gradually shrunk down to the single, efficient, modern form -- before encountering a near-extinction event with the advent of the e-reader.
Place	New York
Publisher	Penguin Books Limited
ISBN	978-0-241-37424-5

# of Pages	230
Date Added	8/4/2022, 5:29:27 AM
Modified	8/11/2022, 6:47:07 AM

- **Tags:**

  - History / Social History
  - Language Arts & Disciplines / General
  - Language Arts & Disciplines / Library & Information Science / Cataloging & Classification
  - Language Arts & Disciplines / Library & Information Science / General
  - Language Arts & Disciplines / Linguistics / Historical & Comparative
  - Literary Criticism / Reference
  - Reference / Bibliographies & Indexes
  - Reference / General
  - SOTAPUB_Index_Chapter

Attachments

  - Google Books Link

- # Indexing Biographies and Other Stories of Human Lives

Item Type	Book
Author	Hazel K. Bell
Abstract	Stories of human lives can be fascinating but frequently difficult to index well. The new, updated fourth edition of Hazel K. Bell's Indexing Biographies is a valuable guide to the points for consideration when indexing life histories, biographies, autobiographies, letters and other narrative texts. Topics include the indexing of fiction, analysis of the text before indexing, names and their various forms, appropriate language choice for index entries, impartiality of the indexer, and how to treat main characters (through appropriate subheading structure) and minor characters (where strings of locators are sometimes unavoidable). The book also discusses more technical matters of index layout, presentation and arrangement of entries, such as how to judge whether alphabetical, chronological, page order or thematic grouping is most appropriate for the text. Examples of good practice and outstanding indexes are provided throughout. Lists of useful reference works and relevant articles from The Indexer journal are also suggested. There is, of course, a comprehensive index. Indexing Biographies contains fine advice on best indexing practices for book indexers, trainee indexers, authors, publishers and all lovers of life histories. It is an excellent overview of the complex, important and rewarding task of indexing such material.
Date	2020-02-26
Language	en
Library Catalog	Google Books
Extra	Google-Books-ID: o4_UDwAAQBAJ
Publisher	Oxford University Press
ISBN	978-1-78962-162-4
# of Pages	150

Date Added	8/11/2022, 2:36:59 PM
Modified	8/11/2022, 2:36:59 PM

- ## Tags:

  - Reference / Bibliographies & Indexes
  - Language Arts & Disciplines / Editing & Proofreading
  - Social Science / Reference

Attachments

  - Google Books Link

- # Indexing fees

Item Type	Web Page
Author	Society of Indexers
Abstract	Indexing fees. Recommended rates.
Language	en-GB
URL	https://www.indexers.org.uk/find-an-indexer/fees-for-indexing
Accessed	8/22/2022, 3:46:47 PM

Extra	WFZ UK rates 2021.
Date Added	8/22/2022, 3:46:47 PM
Modified	8/22/2022, 3:50:17 PM

- ## Attachments

  - Snapshot

- # Indexing: concepts and theory

Item Type	Journal Article
Author	Birger Hjørland
Date	2018
Short Title	Indexing
Library Catalog	Google Scholar
Extra	Publisher: Nomos Verlagsgesellschaft mbH und Co KG
Volume	45
Pages	609
Publication	Knowledge Organization: KO

Issue	7
Date Added	8/11/2022, 7:33:35 AM
Modified	12/4/2022, 11:55:41 PM

- ## Tags:

  - #nosource
  - ⊖ No DOI found

- # Infrastructure: China's Approach to

Item Type	Journal Article
Author	Jeffery Ding
Abstract	Scholars connect China's technology policy to government interventions that target particular industrial sectors. But not all sectors are created equal. Relying on evidence from China's Artificial Intelligence (AI) policies, this paper develops a framework for assessing China's approach toward promoting a technological domain that permeates across many industrial sectors: general-purpose technologies. It shows that China's AI strategy diverges from expectations derived from typical characterizations of China's industrial policy, which stress an emphasis on self-sufficiency, support for a limited number of national champions, and the essential role of military investment and demand for progress in dual-use domains.
Date	2022
Language	en
Library Catalog	Zotero

Extra	WFZ Important corrective to widespread ideas that China's AI policy is dangerously sectoral in focus. Instead, it is dangerously general-purpose in focus! ;-)
Date Added	12/19/2022, 2:47:56 PM
Modified	12/20/2022, 3:03:11 PM

- ## Tags:

  - ⊖ No DOI found

Attachments

- Ding - 2022 - Infrastructure China's Approach to.pdf

- # Inside Book Twitter's Final(?) Days

Item Type	Web Page
Abstract	"It feels like the castle we made is being swept off the table by a billionaire's tantrum," one writer says. Here, insiders tell Esquire how book publishing will change if Twitter goes under.
Date	2023-01-27T11:00:00Z
Language	en-us
Short Title	Inside Book Twitter's Final(?

URL	https://www.esquire.com/entertainment/books/a42638589/book-twitter-end/
Accessed	1/28/2023, 12:55:03 PM
Extra	WFZ We've likely passed peak Book Twitter. Next up: peak Book GPT.
Website Title	Esquire
Date Added	1/28/2023, 12:55:03 PM
Modified	1/29/2023, 1:53:10 PM

- ## Attachments

  - ○ Snapshot

- # InstructZero: Efficient Instruction Optimization for Black-Box Large Language Models

Item Type	Preprint
Author	Lichang Chen
Author	Jiuhai Chen
Author	Tom Goldstein
Author	Heng Huang

Author	Tianyi Zhou

Abstract	Large language models~(LLMs) are instruction followers, but it can be challenging to find the best instruction for different situations, especially for black-box LLMs on which backpropagation is forbidden. Instead of directly optimizing the discrete instruction, we optimize a low-dimensional soft prompt applied to an open-source LLM to generate the instruction for the black-box LLM. On each iteration of the proposed method, which we call InstructZero, a soft prompt is converted into an instruction using the open-source LLM, which is then submitted to the black-box LLM for zero-shot evaluation, and the performance is sent to Bayesian optimization to produce new soft prompts improving the zero-shot performance. We evaluate InstructZero on different combinations of open-source LLMs and APIs including Vicuna and ChatGPT. Our results show that InstructZero outperforms SOTA auto-instruction methods across a variety of downstream tasks. Our code and data are publicly available at https://github.com/Lichang-Chen/InstructZero.

Modified	6/6/2023, 1:52:28 PM

- # Tags:

  - Computer Science - Artificial Intelligence

Notes:

- Comment: 15 pages; 9 figures; Our code is available at https://lichang-chen.github.io/InstructZero/

Attachments

- arXiv Fulltext PDF
- arXiv.org Snapshot

- # Intelligently Extract Text & Data with OCR - Amazon Textract Pricing - Amazon Web Services

Item Type	Web Page
Abstract	Pay only for what you use with Amazon Textract, a machine learning (ML) service that uses optical character recognition (OCR) to automatically extract text, handwriting, and data from scanned PDF documents, forms, and tables.
Date	2022

Language	en-US
URL	https://aws.amazon.com/textract/pricing/
Accessed	9/18/2022, 9:57:37 PM
Extra	WFZ $1.50 for 1000 pages.
Website Title	Amazon Web Services, Inc.
Date Added	9/18/2022, 9:57:37 PM
Modified	9/23/2022, 3:25:14 PM

- ## **Attachments**

  - ○ Snapshot

- # Into the latent space

Item Type	Journal Article
Abstract	Generative deep learning can produce artificial, natural-looking images and other data, which has many promising applications in research — and in art. But the wide availability of generative models poses a challenge for society, which needs tools and best practices to distinguish between real and synthetic data.
Date	2020-03
Language	en

- # Tags:

- o Engineering
- o general

Attachments

- o Full Text PDF
- o Snapshot

- # Introducing Custom instructions! This feature lets you give ChatGPT any custom requests or context which you'd like applied to every conversation. Custom instructions are currently available to Plus users, and we plan to roll out to all users soon! https://t.co/fVIM9GeYk2 Here...

Item Type	Forum Post
Author	OpenAI [@OpenAI]
Date	2023-07-20T17:27Z
Language	en
URL	https://twitter.com/OpenAI/status/1682079674342735872

Accessed	7/20/2023, 11:26:51 PM
Forum/Listserv Title	Twitter
Post Type	Tweet
Date Added	7/20/2023, 11:26:51 PM
Modified	7/20/2023, 11:26:51 PM

- ## Notes:

  - - Novel writing: Using character sheets to help ChatGPT maintain consistent understanding of story characters in ongoing interactions. - Response format: Instructing ChatGPT to consistently output code updates in a unified format. - Writing style personalization: Applying the same voice and style as provided emails to all future email writing request

  - "Introducing Custom instructions! This feature lets you give ChatGPT any custom requests or context which you'd like applied to every conversation. Custom instructions are currently available to Plus users, and we plan to roll out to all users soon! https://t.co/fVIM9GeYk2 Here…"

Attachments

- o Link
- o Snapshot

Introducing Superalignment

Item Type	Web Page
Abstract	We need scientific and technical breakthroughs to steer and control AI systems much smarter than us. To solve this problem within four years, we're starting a new team, co-led by Ilya Sutskever and Jan Leike, and dedicating 20% of the compute we've secured to date to this effort. We're looking for excellent ML researchers and engineers to join us.
Language	en-US
URL	https://openai.com/blog/introducing-superalignment
Accessed	7/5/2023, 1:51:28 PM
Date Added	7/5/2023, 1:51:28 PM
Modified	7/5/2023, 1:51:28 PM

Attachments

- o Snapshot

Invasive Diffusion: How one unwilling illustrator found herself turned into an AI model

Item Type	Blog Post
Author	Andy Baio
Abstract	How does it feel to be turned into an AI image model? To find out, I opened a door to the multiverse and interviewed the creator and unwilling subject of a controversial DreamBooth model.
Date	2022-11-01T15:08:01+00:00
Language	en-US
Short Title	Invasive Diffusion
URL	https://waxy.org/2022/11/invasive-diffusion-how-one-unwilling-illustrator-found-herself-turned-into-an-ai-model/
Accessed	11/2/2022, 9:05:38 AM
Blog Title	Waxy.org
Date Added	11/2/2022, 9:05:38 AM
Modified	11/2/2022, 9:05:42 AM

- # Notes:

  - The result felt like I opened a door to the multiverse, like remaking that scene from Everything Everywhere All at Once, but with me instead of Michelle Yeoh.

Attachments

- Snapshot

Inventors Must Be Human, Federal Circuit Rules in Blow to AI (1)

Item Type	Web Page
Author	Samantha Handler
Abstract	Computer scientist Stephen Thaler was dealt another blow in his battle for artificial intelligence machines to be recognized as inventors on patents, after the nation's top patent court found that inventors must be humans.
Date	2022-08-05
Language	en
URL	https://news.bloomberglaw.com/ip-law/only-humans-not-ai-qualify-as-inventors-federal-circuit-rules
Accessed	8/7/2022, 11:53:35 PM
Extra	WFZ Article summarizes Thaler v. Vidal with good background.
Website Title	Bloomberg Law
Date Added	8/7/2022, 11:53:35 PM

> **Modified**   8/8/2022, 3:29:27 AM

- ## **Attachments**

  - Snapshot

- # Investigations in unsupervised back-of-the-book indexing

Item Type	Conference Paper
Author	Andras Csomai
Author	Rada Mihalcea
Date	2007
Extra	WFZ The most heavily cited paper on the topic.
Proceedings Title	FLAIRS conference
Date Added	8/22/2022, 3:28:33 PM
Modified	12/4/2022, 11:55:42 PM

- ## **Tags:**

  - #nosource
  - ⊖ No DOI found

• Judging a Book By its Cover

Item Type	Preprint
Author	Brian Kenji Iwana
Author	Syed Tahseen Raza Rizvi
Author	Sheraz Ahmed
Author	Andreas Dengel
Author	Seiichi Uchida

Abstract Book covers communicate information to potential readers, but can that same information be learned by computers? We propose using a deep Convolutional Neural Network (CNN) to predict the genre of a book based on the visual clues provided by its cover. The purpose of this research is to investigate whether relationships between books and their covers can be learned. However, determining the genre of a book is a difficult task because covers can be ambiguous and genres can be overarching. Despite this, we show that a CNN can extract features and learn underlying design rules set by the designer to define a genre. Using machine learning, we can bring the large amount of resources available to the book cover design process. In addition, we present a new challenging dataset that can be used for many pattern recognition tasks.

Date	2017-10-12
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/1610.09204
Accessed	8/6/2022, 2:03:53 AM

- **Tags:**

  - Computer Science - Computer Vision and Pattern Recognition

Notes:

  - Comment: 6 pages, 9 figures

Attachments

  - arXiv Fulltext PDF
  - arXiv.org Snapshot

- K-Means Clustering — H2O 3.38.0.3 documentation

Item Type	Web Page
URL	https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/k-means.html#estimating-k-in-k-means
Accessed	1/2/2023, 1:29:38 AM
Extra	WFZ Text clustering for book
Date Added	1/2/2023, 1:29:38 AM
Modified	1/2/2023, 4:59:54 PM

- # Attachments

  - K-Means Clustering — H2O 3.38.0.3 documentation

- # KAIST LK Lab - Home

Item Type	Web Page
URL	https://kaistai.github.io/FLASK/
Accessed	7/21/2023, 11:03:31 PM
Date Added	7/21/2023, 11:03:31 PM
Modified	7/21/2023, 11:03:31 PM

- **Attachments**

  - KAIST LK Lab - Home

- # LaMemo: Language Modeling with Look-Ahead Memory

Item Type	Preprint
Author	Haozhe Ji
Author	Rongsheng Zhang
Author	Zhenyu Yang
Author	Zhipeng Hu
Author	Minlie Huang
Abstract	Although Transformers with fully connected self-attentions are powerful to model long-term dependencies, they are struggling to scale to long texts with thousands of words in language modeling. One of the solutions is to equip the model with a recurrence memory. However, existing approaches directly reuse hidden states from the previous segment that encodes contexts in a uni-directional way. As a result, this prohibits the memory to dynamically interact with the current context that provides up-to-date information for token prediction. To remedy this issue, we propose Look-Ahead Memory (LaMemo) that enhances the recurrence memory by incrementally attending to the right-side tokens, and interpolating with the old memory states to maintain long-term information in the history. LaMemo embraces bi-directional attention and segment recurrence with an additional computation overhead only linearly proportional to the memory length. Experiments on widely used language modeling benchmarks demonstrate its superiority over the baselines equipped with different types of memory.

Date	2022-04-26
Short Title	LaMemo
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2204.07341
Accessed	8/9/2022, 11:04:20 PM
Extra	WFZ Chinese government seems to be funding a lot of longform generation research; NSFC Key project with No. 61936010 comes up frequently. I am beginning to think there is a strategic purpose to influence the global cultural and information environment.
DOI	10.48550/arXiv.2204.07341
Repository	arXiv
Archive ID	arXiv:2204.07341
Date Added	8/9/2022, 11:04:20 PM
Modified	8/9/2022, 11:28:05 PM

- # Tags:

  - Computer Science - Computation and Language

Notes:

- o Comment: Accepted by NAACL 2022

Attachments

- o **arXiv Fulltext PDF**
- o **arXiv.org Snapshot**

• Language Models (Mostly) Know What They Know

Item Type	Preprint
Author	Saurav Kadavath
Author	Tom Conerly
Author	Amanda Askell
Author	Tom Henighan
Author	Dawn Drain
Author	Ethan Perez
Author	Nicholas Schiefer
Author	Zac Hatfield-Dodds
Author	Nova DasSarma
Author	Eli Tran-Johnson

Author   Scott Johnston

Author   Sheer El-Showk

Author   Andy Jones

Author   Nelson Elhage

Author   Tristan Hume

Author   Anna Chen

Author   Yuntao Bai

Author   Sam Bowman

Author   Stanislav Fort

Author   Deep Ganguli

Author   Danny Hernandez

Author   Josh Jacobson

Author   Jackson Kernion

Author   Shauna Kravec

Author   Liane Lovitt

Author   Kamal Ndousse

Author   Catherine Olsson

Author   Sam Ringer

Author   Dario Amodei

Author   Tom Brown

Author   Jack Clark

Author   Nicholas Joseph

Author   Ben Mann

Author   Sam McCandlish

Author   Chris Olah

Author   Jared Kaplan

Abstract   We study whether language models can evaluate the validity of their own claims and predict which questions they will be able to answer correctly. We first show that larger models are well-calibrated on diverse multiple choice and true/false questions when they are provided in the right format. Thus we can approach self-evaluation on open-ended sampling tasks by asking models to first propose answers, and then to evaluate the probability "P(True)" that their answers are correct. We find encouraging performance, calibration, and scaling for P(True) on a diverse array of tasks. Performance at self-evaluation further improves when we allow models to consider many of their own samples before predicting the validity of one specific possibility. Next, we investigate whether models can be trained to predict "P(IK)", the probability that "I know" the answer to a question, without reference to any particular proposed answer. Models perform well at predicting P(IK) and partially generalize across tasks, though they struggle with calibration of P(IK) on new tasks. The predicted P(IK) probabilities also increase appropriately in the presence of relevant source materials in the context, and in the presence of hints towards the solution of mathematical word problems. We hope these observations lay the groundwork for training more honest models, and for investigating how honesty generalizes to cases where models are trained on objectives other than the imitation of human writing.

- ## Tags:

  - Computer Science - Computation and Language
  - Computer Science - Machine Learning
  - Computer Science - Artificial Intelligence

Notes:

- Comment: 23+17 pages; refs added, typos fixed

Attachments

- arXiv Fulltext PDF
- arXiv.org Snapshot

Language-aware Domain Generalization Network for Cross-Scene Hyperspectral Image Classification

Item Type	Preprint
Author	Yuxiang Zhang
Author	Mengmeng Zhang
Author	Wei Li
Author	Shuai Wang
Author	Ran Tao
Abstract	Text information including extensive prior knowledge about land cover classes has been ignored in hyperspectral image classification (HSI) tasks. It is necessary to explore the effectiveness of linguistic mode in assisting HSI classification. In addition, the large-scale pre-training image-text foundation models have demonstrated great performance in a variety of downstream applications, including zero-shot transfer. However, most domain generalization methods have never addressed mining linguistic modal knowledge to improve the generalization performance of model. To compensate for the inadequacies

listed above, a Language-aware Domain Generalization Network (LDGnet) is proposed to learn cross-domain invariant representation from cross-domain shared prior knowledge. The proposed method only trains on the source domain (SD) and then transfers the model to the target domain (TD). The dual-stream architecture including image encoder and text encoder is used to extract visual and linguistic features, in which coarse-grained and fine-grained text representations are designed to extract two levels of linguistic features. Furthermore, linguistic features are used as cross-domain shared semantic space, and visual-linguistic alignment is completed by supervised contrastive learning in semantic space. Extensive experiments on three datasets demonstrate the superiority of the proposed method when compared with state-of-the-art techniques.

Date	2022-09-06
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2209.02700
Accessed	9/13/2022, 4:01:00 AM
Extra	arXiv:2209.02700 [cs]
DOI	10.48550/arXiv.2209.02700
Repository	arXiv
Archive ID	arXiv:2209.02700
Date Added	9/13/2022, 4:01:00 AM
Modified	9/13/2022, 4:01:02 AM

- ## Tags:

- o Computer Science - Computer Vision and Pattern Recognition

Notes:

- o Comment: arXiv admin note: substantial text overlap with arXiv:2209.01634

Attachments

- o arXiv Fulltext PDF
- o arXiv.org Snapshot

Large Language Models Can Self-Improve

Item Type	Preprint
Author	Jiaxin Huang
Author	Shixiang Shane Gu
Author	Le Hou
Author	Yuexin Wu
Author	Xuezhi Wang
Author	Hongkun Yu

Author	Jiawei Han

| **Abstract** | Large Language Models (LLMs) have achieved excellent performances in various tasks. However, fine-tuning an LLM requires extensive supervision. Human, on the other hand, may improve their reasoning abilities by self-thinking without external inputs. In this work, we demonstrate that an LLM is also capable of self-improving with only unlabeled datasets. We use a pre-trained LLM to generate "high-confidence" rationale-augmented answers for unlabeled questions using Chain-of-Thought prompting and self-consistency, and fine-tune the LLM using those self-generated solutions as target outputs. We show that our approach improves the general reasoning ability of a 540B-parameter LLM (74.4%->82.1% on GSM8K, 78.2%->83.0% on DROP, 90.0%->94.4% on OpenBookQA, and 63.4%->67.9% on ANLI-A3) and achieves state-of-the-art-level performance, without any ground truth label. We conduct ablation studies and show that fine-tuning on reasoning is critical for self-improvement. |

Date	2022-10-20
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2210.11610
Accessed	10/24/2022, 5:52:05 AM
Extra	WFZ Human Book Authors (HBAs) Can Self-Improve -- but it is very challenging! Professional authors must be able to edit their own work ruthlessly, and they must also be able to accept editing.
DOI	10.48550/arXiv.2210.11610
Repository	arXiv
Archive ID	arXiv:2210.11610
Date Added	10/24/2022, 5:52:05 AM

Modified	10/24/2022, 5:55:32 AM

- ## Tags:

  - Computer Science - Computation and Language

Attachments

  - **arXiv Fulltext PDF**
  - **arXiv.org Snapshot**

- # Let us serve you, but don't bring us down | Internet Archive Blogs

Item Type	Blog Post
Date	2023-05-29
Language	en-US
URL	https://blog.archive.org/2023/05/29/let-us-serve-you-but-dont-bring-us-down/
Accessed	6/17/2023, 6:20:10 PM
Date Added	6/17/2023, 6:20:10 PM
Modified	6/17/2023, 6:20:10 PM

- # Letting machine learning choose the right font for everyone

Item Type	Web Page
Author	Zoya Bylinskii
Abstract	Personalized Font Recommendations: Combining ML and Typographic Guidelines to Optimize Readability
URL	https://blog.adobe.com/en/publish/2022/06/23/letting-machine-learning-choose-the-right-font-for-everyone
Accessed	6/27/2023, 6:56:57 PM
Website Title	Adobe Blog
Date Added	6/27/2023, 6:56:57 PM
Modified	6/27/2023, 6:56:57 PM

- # **Attachments**

  - Snapshot

- # Little Red Riding Hood Goes Around the Globe:Crosslingual Story Planning

and Generation with Large Language Models

Item Type	Preprint
Author	Evgeniia Razumovskaia
Author	Joshua Maynez
Author	Annie Louis
Author	Mirella Lapata
Author	Shashi Narayan
Abstract	We consider the problem of automatically generating stories in multiple languages. Compared to prior work in monolingual story generation, crosslingual story generation allows for more universal research on story planning. We propose to use Prompting Large Language Models with Plans to study which plan is optimal for story generation. We consider 4 types of plans and systematically analyse how the outputs differ for different planning strategies. The study demonstrates that formulating the plans as question-answer pairs leads to more coherent generated stories while the plan gives more control to the story creators.
Date	2022-12-20
Language	en
Short Title	Little Red Riding Hood Goes Around the Globe
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2212.10471

Accessed	12/28/2022, 6:07:07 PM
Extra	WFZ A good glimpse into what Google is doing in story generation. Plan + Prompts. In the future, books will be written in a language-agnostic way and released everywhere on a nearly concurrent basis, just like other creative goods such as music and movies.
Repository	arXiv
Archive ID	arXiv:2212.10471
Date Added	12/28/2022, 6:07:09 PM
Modified	12/28/2022, 6:10:05 PM

- **Tags:**

  - Computer Science - Computation and Language

Attachments

  - Razumovskaia et al. - 2022 - Little Red Riding Hood Goes Around the GlobeCross.pdf

- # LLaMA: Open and Efficient Foundation Language Models

Item Type	Preprint
Author	Hugo Touvron

Author   Thibaut Lavril

Author   Gautier Izacard

Author   Xavier Martinet

Author   Marie-Anne Lachaux

Author   Timothée Lacroix

Author   Baptiste Rozière

Author   Naman Goyal

Author   Eric Hambro

Author   Faisal Azhar

Author   Aurelien Rodriguez

Author   Armand Joulin

Author   Edouard Grave

Author   Guillaume Lample

Abstract   We introduce LLaMA, a collection of foundation language models ranging from 7B to 65B parameters. We train our models on trillions of tokens, and show that it is possible to train state-of-the-art models using publicly available datasets exclusively, without resorting to proprietary and inaccessible datasets. In particular, LLaMA-13B outperforms GPT-3 (175B) on most benchmarks, and LLaMA-65B is competitive with the best models, Chinchilla-70B and PaLM-540B. We release all our models to the research community.

Date   2023-02-27

Short Title	LLaMA
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2302.13971
Accessed	5/7/2023, 1:01:20 AM
Extra	arXiv:2302.13971 [cs] version: 1
DOI	10.48550/arXiv.2302.13971
Repository	arXiv
Archive ID	arXiv:2302.13971
Date Added	5/7/2023, 1:01:20 AM
Modified	5/7/2023, 1:01:20 AM

- **Tags:**

  - Computer Science - Computation and Language

Attachments

  - arXiv Fulltext PDF
  - arXiv.org Snapshot

- # LlamaIndex

Item Type	Software
Programmer	Jerry Liu
Abstract	LlamaIndex (GPT Index) is a data framework for your LLM applications
Date	2022-11
Library Catalog	GitHub
URL	https://github.com/jerryjliu/llama_index
Accessed	7/23/2023, 7:17:14 PM
Rights	MIT
Extra	DOI: 10.5281/zenodo.1234
Prog. Language	Python
Date Added	7/23/2023, 7:17:14 PM
Modified	7/23/2023, 7:17:14 PM

• LlamaReaders_ppt.pdf

Item Type	Web Page
Abstract	Shared with Dropbox
Language	en

URL	https://www.dropbox.com/s/6e7e16uyptixc1f/LlamaReaders_ppt.pdf?dl=0
Accessed	6/10/2023, 3:49:46 PM
Extra	WFZ Android app that helps "erratic readers" remember plot & character of book *up until the point where they stopped*. Uses LlamaIndex. Clever.
Website Title	Dropbox
Date Added	6/10/2023, 3:49:46 PM
Modified	6/10/2023, 4:59:16 PM

- # LongNet: Scaling Transformers to 1,000,000,000 Tokens

Item Type	Preprint
Author	Jiayu Ding
Author	Shuming Ma
Author	Li Dong
Author	Xingxing Zhang
Author	Shaohan Huang
Author	Wenhui Wang

Author	Furu Wei
Abstract	Scaling sequence length has become a critical demand in the era of large language models. However, existing methods struggle with either computational complexity or model expressivity, rendering the maximum sequence length restricted. In this work, we introduce LongNet, a Transformer variant that can scale sequence length to more than 1 billion tokens, without sacrificing the performance on shorter sequences. Specifically, we propose dilated attention, which expands the attentive field exponentially as the distance grows. LongNet has significant advantages: 1) it has a linear computation complexity and a logarithm dependency between tokens; 2) it can be served as a distributed trainer for extremely long sequences; 3) its dilated attention is a drop-in replacement for standard attention, which can be seamlessly integrated with the existing Transformer-based optimization. Experiments results demonstrate that LongNet yields strong performance on both long-sequence modeling and general language tasks. Our work opens up new possibilities for modeling very long sequences, e.g., treating a whole corpus or even the entire Internet as a sequence.
Date	2023-07-05
Short Title	LongNet
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2307.02486
Accessed	7/5/2023, 11:38:32 PM
Extra	arXiv:2307.02486 [cs]
DOI	10.48550/arXiv.2307.02486
Repository	arXiv
Archive ID	arXiv:2307.02486

Date Added	7/5/2023, 11:38:32 PM
Modified	7/5/2023, 11:38:39 PM

- ## Tags:
  - Computer Science - Computation and Language
  - Computer Science - Machine Learning

- ## Notes:
  - Comment: Work in progress

- ## Attachments
  - arXiv Fulltext PDF
  - arXiv.org Snapshot

- # Machine Learning Model Sizes and the Parameter Gap

Item Type	Journal Article
Author	Pablo Villalobos
Author	Jaime Sevilla
Author	Tamay Besiroglu

Author	Lennart Heim
Author	Anson Ho
Author	Marius Hobbhahn

Abstract We study trends in model size of notable machine learning systems over time using a curated dataset. From 1950 to 2018, model size in language models increased steadily by seven orders of magnitude. The trend then accelerated, with model size increasing by another five orders of magnitude in just 4 years from 2018 to 2022. Vision models grew at a more constant pace, totaling 7 orders of magnitude of growth between 1950 and 2022. We also identify that, since 2020, there have been many language models below 20B parameters, many models above 70B parameters, but a scarcity of models in the 20-70B parameter range. We refer to that scarcity as the parameter gap. We provide some stylized facts about the parameter gap and propose a few hypotheses to explain it. The explanations we favor are: (a) increasing model size beyond 20B parameters requires adopting different parallelism techniques, which makes mid-sized models less cost-effective, (b) GPT-3 was one order of magnitude larger than previous language models, and researchers afterwards primarily experimented with bigger models to outperform it. While these dynamics likely exist, and we believe they play some role in generating the gap, we don't have high confidence that there are no other, more important dynamics at play.

Date	2022
Library Catalog	DOI.org (Datacite)
URL	https://arxiv.org/abs/2207.02852
Accessed	1/24/2023, 8:43:04 PM
Rights	Creative Commons Attribution 4.0 International
Extra	Publisher: arXiv Version Number: 1

- **Tags:**

  - Artificial Intelligence (cs.AI)
  - Computation and Language (cs.CL)
  - FOS: Computer and information sciences
  - Machine Learning (cs.LG)
  - Computers and Society (cs.CY)

- # MACSum: Controllable Summarization with Mixed Attributes

Item Type	Preprint
Author	Yusen Zhang
Author	Yang Liu
Author	Ziyi Yang
Author	Yuwei Fang
Author	Yulong Chen
Author	Dragomir Radev

Author	Chenguang Zhu
Author	Michael Zeng
Author	Rui Zhang

Abstract   Controllable summarization allows users to generate customized summaries with specified attributes. However, due to the lack of designated annotations of controlled summaries, existing works have to craft pseudo datasets by adapting generic summarization benchmarks. Furthermore, most research focuses on controlling single attributes individually (e.g., a short summary or a highly abstractive summary) rather than controlling a mix of attributes together (e.g., a short and highly abstractive summary). In this paper, we propose MACSUM, the first human-annotated summarization dataset for controlling mixed attributes. It contains source texts from two domains, news articles and dialogues, with human-annotated summaries controlled by five designed attributes (Length, Extractiveness, Specificity, Topic, and Speaker). We propose two simple and effective parameter-efficient approaches for the new task of mixed controllable summarization based on hard prompt tuning and soft prefix tuning. Results and analysis demonstrate that hard prompt models yield the best performance on all metrics and human evaluations. However, mixedattribute control is still challenging for summarization tasks. Our dataset and code are available at https://github.com/ psunlpgroup/MACSum.

Date	2022-11-09
Language	en
Short Title	MACSum
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2211.05041
Accessed	11/12/2022, 3:42:59 AM

Extra	WFZ Big names pushing the state of the art. Introducing MACSUM, the first human-annotated summarization dataset for controlling mixed attributes. It contains source texts from two domains, news articles and dialogues, with human-annotated summaries controlled by five designed attributes (Length, Extractiveness, Specificity, Topic, and Speaker). What's missing from the publishing point of view is the motivation: _why_ do you want this particular set of attributes to summarize this particular chunk of text?
Repository	arXiv
Archive ID	arXiv:2211.05041
Date Added	11/12/2022, 3:42:59 AM
Modified	11/12/2022, 5:10:06 AM

- ## Tags:

  - Computer Science - Computation and Language

- ## Notes:

  - Comment: 14 pages, 7 figures

- ## Attachments

  - Zhang et al. - 2022 - MACSum Controllable Summarization with Mixed Attr.pdf

- # Make a search engine in ChatGPT

Item Type	Journal Article
Author	Doug Turnbull
Abstract	Index some documents, provide some queries, ChatGPT will tell you the most relevant documents for those queries
Date	2022-12-04
URL	https://softwaredoug.com/blog/2022/12/04/make-a-search-engine-in-chatgpt.html
Accessed	12/4/2022, 5:23:46 PM
Extra	WFZ Search guru Doug Turnbull (now Reddit) glimpses the future of search as conversation using ChatGPT. Important and insightul.
Publication	Doug Turnbull's Blog
Date Added	12/4/2022, 5:23:46 PM
Modified	12/5/2022, 12:36:56 AM

- ## Attachments

  - ○ Snapshot

- # Max-Planck-Institut für Informatik: Children's texts for commonsense knowledge extraction

Item Type	Web Page
URL	https://www.mpi-inf.mpg.de/children-texts-for-commonsense
Accessed	2/16/2023, 5:23:18 PM
Date Added	2/16/2023, 5:23:18 PM
Modified	2/16/2023, 5:23:23 PM

- **Attachments**

  - Max-Planck-Institut für Informatik: Children's texts for commonsense knowledge extraction

- # Meeting Summarization: A Survey of the State of the Art

Item Type	Conference Paper
Author	L. P. Kumar
Author	Arman Kabiri
Abstract	Information overloading requires the need for summarizers to extract salient informat platforms. The rise of Covid-19 has led people to rely on online communication platfo through the entire meeting transcripts, people can use meeting summarizers to select u this survey, we aim to cover recent meeting summarization techniques. Our survey of summarization. We also provide the performance of each summarizer on a leaderboar future researchers.

Date	16 December 2022
Short Title	Meeting Summarization
Library Catalog	Semantic Scholar
URL	[https://www.semanticscholar.org/paper/Meeting-Summarization%3A-A-Survey-of-th](https://www.semanticscholar.org/paper/Meeting-Summarization%3A-A-Survey-of-the) Kabiri/e29f6d3f72dc57ee08feb04865a490a117a6e270?utm_source=alert_email&utm 0-1&utm_medium=10036602
Accessed	12/22/2022, 12:49:42 PM
Extra	WFZ I had to cite this for the opportunity to joke about "Turn Your Meeting Into A B understanding dialogue these days, and it's important to future directions in long form
Date Added	12/22/2022, 12:49:42 PM
Modified	12/22/2022, 10:46:46 PM

- ## **Notes:**

  - ○ [TLDR] This survey offers a general overview of text summarization along with datasets and evaluation metrics for meeting summarization, and provides the performance of each summarizer on a leaderboard.

Attachments

- ○ Full Text PDF
- ○ Semantic Scholar Link

• Memory for artwork is predictable

Item Type	Journal Article
Author	Trent M. Davis
Author	Wilma A. Bainbridge
Abstract	Viewing art is often seen as a highly personal and subjective experience. However, are there universal factors that make a work of art memorable? We conducted three experiments, where we recorded online memory performance for 4,021 paintings from the Art Institute of Chicago, tested in-person memory after an unconstrained visit to the Art Institute, and obtained abstract attribute measures such as beauty and emotional valence for these pieces. Participants showed significant agreement in their memories both online and in-person, suggesting that pieces have an intrinsic "memorability" based solely on their visual properties that is predictive of memory in a naturalistic museum setting. Importantly, ResMem, a deep learning neural network designed to estimate image memorability, could significantly predict memory both online and in-person based on the images alone, and these predictions could not be explained by other low- or high-level attributes like color, content type, aesthetics, and emotion. A regression comprising ResMem and other stimulus factors could predict as much as half of the variance of in-person memory performance. Further, ResMem could predict the fame of a piece, despite having no cultural or historical knowledge. These results suggest that perceptual features of a painting play a major role in influencing its success, both in memory for a museum visit and in cultural memory over generations.
Date	2023-07-11
Library Catalog	pnas.org (Atypon)
URL	https://www.pnas.org/doi/10.1073/pnas.2302389120
Accessed	7/25/2023, 12:22:06 AM

Extra	WFZ There is science to what makes memorable art, and it is different from, and independent of, what people think of as beauty. Very possibly true for text too.
Volume	120
Pages	e2302389120
Publication	Proceedings of the National Academy of Sciences
DOI	10.1073/pnas.2302389120
Issue	28
Date Added	7/25/2023, 12:22:06 AM
Modified	7/27/2023, 2:23:31 AM

- # Attachments

  - ○ Full Text PDF

- # Meta is putting its latest AI chatbot on the web for the public to talk to - The Verge

Item Type	Web Page
Date	8/5/2022

URL	https://www.theverge.com/2022/8/5/23293281/meta-ai-chatbot-blenderbot-3-web-access-research-safety
Accessed	8/5/2022, 9:55:57 PM
Extra	WFZ Blenderbot makes its research answers traceable, which is important because that raises the bar to where book authors and publishers need it to be.
Date Added	8/5/2022, 9:55:57 PM
Modified	8/6/2022, 12:33:19 AM

- ## Attachments

  - Meta is putting its latest AI chatbot on the web for the public to talk to - The Verge

- # Microsoft Bing Chat (Sydney) – Dr Alan D. Thompson – Life Architect

Item Type	Web Page
URL	https://lifearchitect.ai/bing-chat/
Accessed	2/13/2023, 3:48:33 PM
Date Added	2/13/2023, 3:48:33 PM
Modified	2/13/2023, 3:48:33 PM

- **Attachments**

  o Microsoft Bing Chat (Sydney) – Dr Alan D. Thompson – Life Architect

- # Microsoft empowers Windows, iOS, Android, Mac and Linux developers to reach billions of new customers - Stories

Item Type	Web Page
Author	Microsoft Corporation
Date	2015
URL	https://news.microsoft.com/2015/04/29/microsoft-empowers-windows-ios-android-mac-and-linux-developers-to-reach-billions-of-new-customers/
Accessed	8/24/2022, 11:06:56 PM
Extra	WFZ Most recent readily searchable estimate of number of Office users = 1.2 billion.
Date Added	8/24/2022, 11:06:56 PM
Modified	8/24/2022, 11:08:29 PM

- **Tags:**

- o ETL

Attachments

- o Microsoft empowers Windows, iOS, Android, Mac and Linux developers to reach billions of new customers - Stories

- # Microsoft eyes $10 billion bet on ChatGPT | Semafor

Item Type	Magazine Article
Author	Liz Hoffman
Author	Reid Albergotti
Abstract	The tech giant has been in talks on deal to effectively own almost half of OpenAI, maker of the addictive, humanoid, AI-powered chatbot.
Date	2023-01-10T02:28:10Z
Language	en
URL	https://www.semafor.com/article/01/09/2023/microsoft-eyes-10-billion-bet-on-chatgpt
Accessed	1/10/2023, 11:27:24 AM
Extra	Section: business

- # Attachments

  o ## Snapshot

- # Microsoft in Talks to Acquire 49% of OpenAI

Item Type	Web Page
Author	Eddie Donmez
Abstract	Microsoft would receive 75% of OpenAI's income until it has recovered its initial investment, good terms for the tech giant.
URL	https://www.linkedin.com/posts/activity-7018490472520142848-iEKj
Accessed	1/10/2023, 11:12:48 AM
Extra	WFZ The eyebrow-raisers here are that MSFT would take a 49% stake and receive 75% of OpenAI's income until forever. This reminds me of one of Kevin O'Leary's horrible deals on #SharkTank. No, no, no. Don't do this, OpenAI. All he cares about is selling his stupid wine. Besides, if you have ever seen an episode of any murder mystery ever, you realize that owning 51% of something means that you can lose control if just one moderately large shareholder decides to "flip" and team up with the minority stakeholder.

Date Added	1/10/2023, 11:12:48 AM
Modified	1/10/2023, 11:17:26 AM

- ## Attachments

  - (2) Post | Feed | LinkedIn

- # Mubert-Text-to-Music

Item Type	Software
Abstract	A simple notebook demonstrating prompt-based music generation via Mubert API
Date	2022-10-20T11:44:48Z
Library Catalog	GitHub
URL	https://github.com/MubertAI/Mubert-Text-to-Music/blob/d1262380a58afe93c59480bbc4ef87546bc0c714/Mubert_Text_to_Music.
Accessed	10/20/2022, 7:51:44 AM
Extra	WFZ Demonstrates text-to-music in Google CoLab using Mubert API. Unfortunately Mubert begins at $200/month, but, in principle, any book can now be accompanied by own bespoke, royalty-free mood music.
Company	Mubert: music powered by AI
Prog. Language	Jupyter Notebook

Date Added	10/20/2022, 7:51:44 AM
Modified	12/4/2022, 11:55:51 PM

- **Tags:**

  - #nosource

- # Multilingual Folk Tale Database

Item Type	Web Page
Date	2016
URL	http://www.mftd.org/index.php?action=atu
Accessed	12/3/2022, 10:59:02 AM
Date Added	12/3/2022, 10:59:02 AM
Modified	12/5/2022, 12:21:47 AM

- **Attachments**

  - MFTD

- # Multimodal Machine Learning: A Survey and Taxonomy

Item Type	Preprint
Author	Tadas Baltrušaitis
Author	Chaitanya Ahuja
Author	Louis-Philippe Morency

| **Abstract** | Our experience of the world is multimodal - we see objects, hear sounds, feel texture, smell odors, and taste flavors. Modality refers to the way in which something happens or is experienced and a research problem is characterized as multimodal when it includes multiple such modalities. In order for Artificial Intelligence to make progress in understanding the world around us, it needs to be able to interpret such multimodal signals together. Multimodal machine learning aims to build models that can process and relate information from multiple modalities. It is a vibrant multi-disciplinary field of increasing importance and with extraordinary potential. Instead of focusing on specific multimodal applications, this paper surveys the recent advances in multimodal machine learning itself and presents them in a common taxonomy. We go beyond the typical early and late fusion categorization and identify broader challenges that are faced by multimodal machine learning, namely: representation, translation, alignment, fusion, and co-learning. This new taxonomy will enable researchers to better understand the state of the field and identify directions for future research. |

Date	2017-08-01
Language	en
Short Title	Multimodal Machine Learning
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/1705.09406
Accessed	10/8/2022, 8:21:39 AM

Extra	arXiv:1705.09406 [cs]
Repository	arXiv
Archive ID	arXiv:1705.09406
Date Added	10/8/2022, 8:21:42 AM
Modified	10/8/2022, 8:21:44 AM

- ## Tags:

  - Computer Science - Machine Learning

Attachments

- Baltrušaitis et al. - 2017 - Multimodal Machine Learning A Survey and Taxonomy.pdf

- # Must read: the 100 most cited AI papers in 2022

Item Type	Web Page
Author	Sergi Castella i Sapé
Abstract	Who Is publishing the most Impactful AI research right now? With the breakneck pace of innovation in AI, it is crucial to pick up some signal as soon as possible. No one has the time to read everything, but these 100 papers are sure to bend the road as to where our AI technology is going. The real test of impact of R&D teams is of course how the technology appears in products, and OpenAI shook the world

	by releasing ChatGPT at the end of November 2022, following fast on their March 2022 paper "T
Date	2023-03-02T09:07:35.250Z
Language	en
Short Title	Must read
URL	https://www.zeta-alpha.com/post/must-read-the-100-most-cited-ai-papers-in-2022
Accessed	3/5/2023, 8:01:32 PM
Website Title	Zeta Alpha
Date Added	3/5/2023, 8:01:32 PM
Modified	3/5/2023, 8:01:32 PM

- ## Attachments

  - ○ Snapshot

- # Narrative flow: a formal approach of character network generation for non-linear narratives

Item Type	Journal Article

Author	Coline Métrailler
Language	en
Library Catalog	Zotero
Pages	3
Date Added	6/22/2022, 5:52:23 PM
Modified	6/22/2022, 5:52:23 PM

- ## Tags:

  - ⊖ No DOI found

Attachments

  - Métrailler - Narrative flow a formal approach of character net.pdf

- # NETWORK ANALYSIS OF THE KYIV BYLYNY CYCLE — EAST SLAVIC EPIC NARRATIVES

Item Type	Journal Article
Author	Petro Sarkanych

Author	Nazar Fedorak
Author	Yurij Holovatch
Author	Pádraig Maccarron
Author	Joseph Yose
Author	Ralph Kenna
Abstract	In recent times, the advent of network science permitted new quantitative approaches to literary studies. Here, we bring the Kyiv bylyny cycle into the field — East Slavic epic narratives originating in modern-day Ukraine. By comparing them to other prominent European epics, we identify universal and distinguishing properties of the social networks in bylyny. We analyze community structures and rank most important characters. The method allows to bolster hypotheses from humanities literature — such as the position of Prince Volodymyr — and to generate new ones. We show how the Kyiv cycle of bylyny fits very well with narrative networks from other nations — especially heroic ones. We anticipate that, besides delivering new narratological insights, this study will aid future scholars and interested public to navigate their way through Ukraine's epic story and identify its heroes.
Date	08/2022
Language	en
Library Catalog	DOI.org (Crossref)
URL	https://www.worldscientific.com/doi/10.1142/S0219525922400070
Accessed	12/2/2022, 11:15:37 PM
Extra	WFZ The "bylyny" are epic songs from 9th-11C Kyiv and Ukraine. The main content of the article compares the network graph metrics of bylyny network to other saga networks including the Odyssey, Beowulf, Irish, and Icelandic sagas

(see Table 2 at PDF page 10). It fits well with metrics for these European networks, suggesting that the metrics are robust enough to support reverse-engineering from metrics to content. This article also contains some good pointers into the global literature for folklore.

- **Attachments**

  - Full Text

- # Networks and Knowledge in Roget's Thesaurus

Item Type	Book

Author	Werner Hüllen
Abstract	Abstract. This book epitomizes the history of Roget's Thesaurus as the merger of a dictionary of synonyms and a topical, i.e. non-alphabetical, dictionary. It h
Date	2009/1/01
Language	en
Library Catalog	academic.oup.com
URL	https://academic.oup.com/book/8865
Accessed	8/29/2022, 9:58:14 AM
Extra	DOI: 10.1093/acprof:oso/9780199553235.001.0001 DOI: 10.1093/acprof:oso/9780199553235.001.0001
Date Added	8/29/2022, 9:58:13 AM
Modified	8/29/2022, 9:58:22 AM

- ## Attachments

  - ○ Snapshot

- # Neural Graph Databases

Item Type	Web Page
Author	Michael Galkin

Abstract	A new milestone in graph data management
Date	2023-03-28T14:08:15.825Z
Language	en
URL	https://towardsdatascience.com/neural-graph-databases-cc35c9e1d04f
Accessed	3/28/2023, 7:10:57 PM
Website Title	Medium
Date Added	3/28/2023, 7:10:59 PM
Modified	3/28/2023, 7:10:59 PM

- # NeuroLogic A*esque decoding: Constrained text generation with lookahead heuristics

Item Type	Journal Article
Author	Ximing Lu
Author	Sean Welleck
Author	Peter West
Author	Liwei Jiang

Author   Jungo Kasai

Author   Daniel Khashabi

Author   Ronan Le Bras

Author   Lianhui Qin

Author   Youngjae Yu

Author   Rowan Zellers

Author   Noah A. Smith

Author   Yejin Choi

Date   2022

Extra   WFZ Book publishing is all about constraints--copyright, topic, style, length, format, genre--so constrained text generation is important to AI for book-lovers. I found the section about constraining translation to particular vocabularies especially interesting as I edit and publish in many genres that have exacting terminological requirements. Worth mentioning that this paper relies on Mechanical Turk for evaluating story generations--see Karpsinska et al. 2021. Important to figure out ways to get book people involved.

Volume   abs/2112.08726

Publication   ArXiv

Date Added   8/18/2022, 10:09:29 PM

Modified   12/4/2022, 11:55:42 PM

- **Tags:**

  - #nosource
  - ⊖ No DOI found

- # New GPT-3 model: text-davinci-003

Item Type	Letter
Author	OpenAI
Abstract	It produces higher quality writing. This will help your applications deliver clearer, more engaging, and more compelling content. It can handle more complex instructions, meaning you can get even more creative with how you make use of its capabilities now. It's better at longer form content generation, allowing you to take on tasks that would have previously been too difficult to achieve
Date	2022-11-29
Short Title	davinci-003 announcement
Extra	WFZ Major quality improvements in GPT3. OpenAI says: "It produces higher quality writing. This will help your applications deliver clearer, more engaging, and more compelling content. It can handle more complex instructions, meaning you can get even more creative with how you make use of its capabilities now. It's better at longer form content generation, allowing you to take on tasks that would have previously been too difficult to achieve." More specifics on exactly what has improved would be helpful. Maybe there will be a journal article in future.
Date Added	11/29/2022, 5:24:08 AM
Modified	12/4/2022, 11:56:02 PM

- **Tags:**

  - #nosource

- # New Study: 55% of YA Books Bought by Adults

Item Type	Web Page
Abstract	More than half the consumers of books classified for young adults aren't all that young. According to a new study, fully 55% of buyers of works that publishers designate for kids aged 12 to 17 -- nicknamed YA books -- are 18 or older, with the largest segment aged 30 to 44, which alone accounted for 28% of YA sales. And adults aren't just purchasing for others -- when asked about the intended recipient, they report that 78% of the time they are purchasing books for their own reading.
Language	en
Short Title	New Study
URL	https://www.publishersweekly.com/pw/by-topic/childrens/childrens-industry-news/article/53937-new-study-55-of-ya-books-bought-by-adults.html
Accessed	5/23/2023, 8:04:44 PM
Website Title	PublishersWeekly.com
Date Added	5/23/2023, 8:04:44 PM
Modified	5/23/2023, 8:04:44 PM

- **Attachments**

  ○ Snapshot

- # New York Times drops out of AI coalition | Semafor

Item Type	Web Page
Abstract	The move is a major blow to efforts to Barry Diller's efforts to establish an industry united front against Google and Microsoft.
Date	2023-08-14T00:29:05Z
Language	en
URL	https://www.semafor.com/article/08/13/2023/new-york-times-drops-out-of-ai-coalition
Accessed	8/13/2023, 9:37:52 PM
Extra	Section: media
Date Added	8/13/2023, 9:37:52 PM
Modified	8/13/2023, 9:37:52 PM

- # Notes:

- One person said that the Times had discussed joining the group, but never committed. The decision by the Times makes it more likely that publishers may cut their own separate deals. Earlier this year, the Associated Press announced an agreement to license the AP's archive of news stories to ChatGPT-maker OpenAI for two years

Attachments

- Snapshot

- # News Corp using AI to produce 3,000 Australian local news stories a week

Item Type	Newspaper Article
Author	Amanda Meade
Abstract	The Data Local team uses AI technology to generate stories on weather, fuel prices and traffic reports for hyperlocal mastheads
Date	2023-07-31T15:00:41.000Z
Language	en-GB
Library Catalog	The Guardian
URL	https://www.theguardian.com/media/2023/aug/01/news-corp-ai-chat-gpt-stories

- ## Tags:

  - Technology
  - Artificial intelligence (AI)
  - Australia news
  - Australian media
  - Business
  - Data journalism
  - Media
  - Media business
  - News Corporation
  - Rupert Murdoch

Notes:

  - News Corp Australia is producing 3,000 articles a week using generative artificial intelligence, executive chair Michael Miller has revealed.

Miller told the World News Media Congress in Taipei that a team of four staff use the technology to generate thousands of local stories each week on weather, fuel prices and traffic conditions, according to a report in Mediaweek.

Attachments

- ○ Snapshot

No Language Left Behind: Scaling Human-Centered Machine Translation

Item Type	Journal Article
Author	NLLB Team
Author	Marta R Costa-jussà
Author	James Cross
Author	Onur Çelebi
Author	Maha Elbayad
Author	Kenneth Heafield
Author	Kevin Heffernan

Author   Elahe Kalbassi

Author   Janice Lam

Author   Daniel Licht

Author   Jean Maillard

Author   Anna Sun

Author   Skyler Wang

Author   Guillaume Wenzek

Author   Al Youngblood

Author   Bapi Akula

Author   Loic Barrault

Author   Gabriel Mejia Gonzalez

Author   Prangthip Hansanti

Author   John Hoffman

Author   Semarley Jarrett

Author   Kaushik Ram Sadagopan

Author   Dirk Rowe

Author   Shannon Spruit

Author   Chau Tran

Author   Pierre Andrews

Author   Necip Fazil Ayan

Author   Shruti Bhosale

Author   Sergey Edunov

Author   Angela Fan

Author   Cynthia Gao

Author   Vedanuj Goswami

Author   Francisco Guzmán

Author   Philipp Koehn

Author   Alexandre Mourachko

Author   Christophe Ropers

Author   Safiyyah Saleem

Author   Holger Schwenk

Author   Jeff Wang

Author   Meta Ai

Abstract   Driven by the goal of eradicating language barriers on a global scale, machine translation has solidified itself as a key focus of artificial intelligence research

today. However, such efforts have coalesced around a small subset of languages, leaving behind the vast majority of mostly low-resource languages. What does it take to break the 200 language barrier while ensuring safe, high quality results, all while keeping ethical considerations in mind? In No Language Left Behind, we took on this challenge by first contextualizing the need for low-resource language translation support through exploratory interviews with native speakers. Then, we created datasets and models aimed at narrowing the performance gap between low and high-resource languages. More specifically, we developed a conditional compute model based on Sparsely Gated Mixture of Experts that is trained on data obtained with novel and effective data mining techniques tailored for low-resource languages. We propose multiple architectural and training improvements to counteract overfitting while training on thousands of tasks. Critically, we evaluated the performance of over 40,000 different translation directions using a human-translated benchmark, Flores-200, and combined human evaluation with a novel toxicity benchmark covering all languages in Flores-200 to assess translation safety. Our model achieves an improvement of 44% BLEU relative to the previous state-of-the-art, laying important groundwork towards realizing a universal translation system. Finally, we open source all contributions described in this work, accessible at https://github.com/facebookresearch/fairseq/tree/nllb.

Language	en
Library Catalog	Zotero
Pages	190
Date Added	7/7/2022, 1:01:24 AM
Modified	7/7/2022, 1:01:25 AM

- ## Tags:

  - ⊖ No DOI found

Attachments

- ○ Team et al. - No Language Left Behind Scaling Human-Centered Ma.pdf

Nomic AI

Item Type	Web Page
URL	https://home.nomic.ai/visxwiki
Accessed	11/2/2022, 8:12:47 PM
Date Added	11/2/2022, 8:12:47 PM
Modified	11/2/2022, 8:12:47 PM

Attachments

- ○ Nomic AI

Notes for Multimodality course from CMU

Item Type	Web Page
Author	Chadha, Aman and Jain, Vinija
Date	2022-07-01
URL	https://vinija.ai/multimodal/

Accessed	10/8/2022, 8:15:02 AM
Extra	WFZ Amazon ML engineer reviews CMU course on multimodal. ~10-15 pp wih images, glossary, simple explanations of concepts. Good place to start.
Website Title	Vinjia's AI
Date Added	10/8/2022, 8:15:02 AM
Modified	12/4/2022, 11:55:46 PM

- **Tags:**

  - #nosource

Novel Views on Novels: Embedding Multiple Facets of Long Texts

Item Type	Journal Article
Author	Lasse Kohlmeyer
Author	Tim Repke
Author	Ralf Krestel
Abstract	Novels are one of the longest document types and thus one of the most complex types of texts. Many NLP tasks utilize document embeddings as machine-understandable semantic representations of documents. However, such document embeddings are optimized for short texts, such as sentences or paragraphs. When faced with longer texts, these models either truncate the long text or split it

sequentially into smaller chunks. We show that when applied to a fictional novel, these traditional document embeddings fail to capture all its facets. Complex information, such as time, place, atmosphere, style, and plot is typically not represented adequately. To this end, we propose lib2vec which computes and combines multiple embedding vectors based on various facets. Instead of splitting the text sequentially, lib2vec splits the text semantically based on domain-specific facets. We evaluate the semantic expressiveness using human-assessed book comparisons as well as content-based information retrieval tasks. The results show that our approach outperforms state-of-the-art document embeddings for long texts.

URL	https://www.goodreads.com/
Accessed	10/26/2021, 8:00:00 PM
Extra	ISBN: 0000000000000
Date Added	10/27/2021, 5:30:13 PM
Modified	12/4/2022, 11:55:08 PM

- **Tags:**

  - #nosource
  - ⊖ No DOI found

- # OCR with Tesseract, Amazon Textract, and Google Document AI: a benchmarking experiment

Item Type	Journal Article

Author	Thomas Hegghammer
Abstract	Optical Character Recognition (OCR) can open up understudied historical documents to computational analysis, but the accuracy of OCR software varies. This article reports a benchmarking experiment comparing the performance of Tesseract, Amazon Textract, and Google Document AI on images of English and Arabic text. English-language book scans (n = 322) and Arabic-language article scans (n = 100) were replicated 43 times with different types of artificial noise for a corpus of 18,568 documents, generating 51,304 process requests. Document AI delivered the best results, and the server-based processors (Textract and Document AI) performed substantially better than Tesseract, especially on noisy documents. Accuracy for English was considerably higher than for Arabic. Specifying the relative performance of three leading OCR products and the differential effects of commonly found noise types can help scholars identify better OCR solutions for their research needs. The test materials have been preserved in the openly available "Noisy OCR Dataset" (NOD) for reuse in future benchmarking studies.
Date	2022-05-01
Language	en
Short Title	OCR with Tesseract, Amazon Textract, and Google Document AI
Library Catalog	Springer Link
URL	https://doi.org/10.1007/s42001-021-00149-1
Accessed	9/17/2022, 7:52:02 PM
Extra	WFZ Google
Volume	5
Pages	861-882

- **Tags:**

    - Benchmarking
    - Cloud computing
    - OCR

Attachments

    - Full Text PDF

- # On Capitol Hill, Humans Battle AI Over Authorship | CCC's Velocity of Content Podcast

Item Type	Blog Post

Abstract	As chatbots rise, US Authors Guild lobbies for incentives for human creators to continue working.
Language	en-US
URL	https://velocityofcontentpodcast.com/on-capitol-hill-humans-battle-ai-over-authorship/
Accessed	4/7/2023, 2:54:30 PM
Extra	WFZ Promising start to the article, but goes downhill fast when it becomes a plea for $$$ for the 13,000 members of the Authors Guild. Not likely to be granted.
Blog Title	Velocity of Content | A series of recordings from the Copyright Clearance Center
Date Added	4/7/2023, 2:54:29 PM
Modified	4/8/2023, 6:35:35 PM

- # **Attachments**

  - ## Snapshot

- # On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ⸬⸬⸬

Item Type	Conference Paper
Author	Emily M. Bender

Author	Timnit Gebru
Author	Angelina McMillan-Major
Author	Shmargaret Shmitchell
Date	2021-03-03
Language	en
Short Title	On the Dangers of Stochastic Parrots
Library Catalog	DOI.org (Crossref)
URL	https://dl.acm.org/doi/10.1145/3442188.3445922
Accessed	7/19/2022, 2:25:31 PM
Place	Virtual Event Canada
Publisher	ACM
ISBN	978-1-4503-8309-7
Pages	610-623
Proceedings Title	Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency
Conference Name	FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency
DOI	10.1145/3442188.3445922

Date Added	3/3/2021, 8:03:30 PM
Modified	10/19/2022, 11:06:44 PM

- **Tags:**
  - #nosource

Attachments

  - Full Text

- # One Graph to Rule them All: Using NLP and Graph Neural Networks to analyse Tolkien's Legendarium

Item Type	Journal Article
Author	Vincenzo Perri
Author	Lisi Qarkaxhija
Author	Albin Zehe
Author	Andreas Hotho
Author	Ingo Scholtes
Date	2022

Short Title	One Graph to Rule them All
Library Catalog	Google Scholar
Publication	arXiv preprint arXiv:2210.07871
Date Added	10/19/2022, 10:29:59 PM
Modified	10/19/2022, 10:30:02 PM

- ## Tags:

  - ⊖ No DOI found

Attachments

  - Full Text
  - Snapshot

- # OpenAI API

Item Type	Web Page
Abstract	An API for accessing new AI models developed by OpenAI
Language	en
URL	https://platform.openai.com

Accessed	1/31/2023, 6:40:49 PM
Date Added	1/31/2023, 6:40:49 PM
Modified	1/31/2023, 6:40:53 PM

- ## Attachments

  - Snapshot

- # OpenAI Used Kenyan Workers on Less Than $2 Per Hour: Exclusive | Time

Item Type	Web Page
URL	https://time.com/6247678/openai-chatgpt-kenya-workers/?utm_source=roundup&utm_campaign=20230105
Accessed	1/18/2023, 10:49:11 PM
Extra	WFZ Outstanding reporting. Great detail about the sweatshop work that goes into AI training. More than just a pause for thought for anyone who relies upon and enjoys this technology.
Date Added	1/18/2023, 10:49:11 PM
Modified	1/20/2023, 6:49:52 AM

- # OpenAI's attempts to watermark AI text hit limits

Item Type	Blog Post
Author	Kyle Wiggers
Abstract	OpenAI is working on a way to watermark text, according to a guest researcher working at the lab. But it's limited in key ways.
Date	2022-12-10T13:15:11+00:00
Language	en-US
URL	https://techcrunch.com/2022/12/10/openais-attempts-to-watermark-ai-text-hit-limits/
Accessed	12/10/2022, 9:12:53 AM
Blog Title	TechCrunch
Date Added	12/10/2022, 9:12:53 AM
Modified	12/15/2022, 8:48:08 AM

- # Tags:
  - Watermarking

Notes:

- In a lecture at the University of Austin, computer science professor Scott Aaronson, currently a guest researcher at OpenAI, revealed that OpenAI is developing a tool for "statistically watermarking the outputs of a text [AI system]." Whenever a system — say, ChatGPT — generates text, the tool would embed an "unnoticeable secret signal" indicating where the text came from.

Attachments

- Snapshot

OpenAI's GPT-3 produced some crazy color names - Protocol

Item Type	Web Page
Author	Kate Kaye
Abstract	Turns out, sometimes it's OK when GPT-3's natural language AI models make up stuff.
Language	en
URL	https://www.protocol.com/bulletins/ai-text-generator-color-wheel
Accessed	1/29/2023, 4:38:30 PM

Date Added	1/29/2023, 4:38:30 PM
Modified	1/29/2023, 4:38:30 PM

- **Attachments**

  - Snapshot

- # OpenJourney: Midjourney, but Open Source | Hacker News

Item Type	Web Page
URL	https://news.ycombinator.com/item?id=34522311
Accessed	1/26/2023, 8:20:39 PM
Date Added	1/26/2023, 8:20:41 PM
Modified	1/26/2023, 8:20:41 PM

- # Opinion | Let ChatGPT and Quillbot Write Your Book

Item Type	Newspaper Article
Author	Peter Funt

Abstract	Maybe they'll produce a bestseller, though they probably won't.
Date	2023-08-16T22:18:00Z
Language	en-US
Library Catalog	www.wsj.com
URL	https://www.wsj.com/articles/chatgpt-quillbot-write-your-book-amazon-self-publishing-ai-artificial-intelligence-ebook-kindle-210e3ab6
Accessed	8/22/2023, 9:43:27 PM
Section	Opinion
Publication	Wall Street Journal
ISSN	0099-9660
Date Added	8/22/2023, 9:43:27 PM
Modified	8/22/2023, 9:43:38 PM

- # Tags:

  - sciences
  - literature
  - publishing
  - Technology
  - machine learning
  - artificial intelligence

- books
- ai
- Artificial Intelligence Technologies
- Artificial Intelligence/Machine Learning
- arts
- Arts/Entertainment
- Book Publishing
- book sales
- C&E Executive News Filter
- commentaries
- Commentaries/Opinions
- Computer Science
- Content Types
- Domestic Politics
- e-books
- E-commerce
- ebooks
- entertainment
- Etailing
- Factiva Filters
- general news
- humanities
- international relations
- media
- Media/Entertainment
- Online Service Providers
- Opinion
- opinions
- political
- Political/General News

- politics
- Politics/International Relations
- printing
- Printing/Publishing
- Retail
- Retail/Wholesale
- Sciences/Humanities
- self-publishing
- SYND
- wholesale
- writing
- WSJ-PRO-WSJ.com

Notes:

- You'll have a carton of unsold copies in your garage as a reminder that while writing books these days is simpler than ever, selling them is practically impossible.

Attachments

- Snapshot

Opinion | This Film Does Not Exist

Item Type	Newspaper Article
Author	Frank Pavich

Abstract	How artificial intelligence can reimagine art from our past and influence our future.
Date	2023-01-13
Language	en-US
Library Catalog	NYTimes.com
URL	https://www.nytimes.com/interactive/2023/01/13/opinion/jodorowsky-dune-ai-tron.html
Accessed	1/14/2023, 12:53:57 PM
Section	Opinion
Publication	The New York Times
ISSN	0362-4331
Date Added	1/14/2023, 12:53:57 PM
Modified	1/14/2023, 12:53:57 PM

- # Tags:

  - Artificial Intelligence
  - Jodorowsky, Alejandro

Attachments

Opinion | Watch Out for the Fake Tom Cruise

Item Type	Newspaper Article
Author	Maureen Dowd
Abstract	Hollywood is about to get even more phony!
Date	2023-07-15
Language	en-US
Library Catalog	NYTimes.com
URL	https://www.nytimes.com/2023/07/15/opinion/writers-actors-strike.html
Accessed	7/15/2023, 10:45:49 PM
Extra	WFZ The money quote here is from Jaron Lanier: "They might say, 'Make me a movie that's similar to Tom Cruise's "Mission: Impossible." However, make sure that none of the synthetic actors can be mistaken for known actors and make sure that we're not going to get sued, but let's go right up to the line.' That's not quite feasible today, but I don't see any reason why it wouldn't be. It's just math. And we can do it." Same for books, right?
Section	Opinion
Publication	The New York Times

- ## Tags:

  - Artificial Intelligence
  - Hollywood (Calif)
  - Lanier, Jaron
  - SAG-AFTRA
  - Silverman, Sarah
  - Simone (Movie)
  - Strikes
  - Writers Guild of America

Notes:

  - As Lanier said, "They might say, 'Make me a movie that's similar to Tom Cruise's "Mission: Impossible." However, make sure that none of the synthetic actors can be mistaken for known actors and make sure that we're not going to get sued, but let's go right up to the line.' That's not quite feasible today, but I don't see any reason why it wouldn't be. It's just math. And we can do it."

Attachments

- ## Opt-Me-Out From Diffusion: This AI Model Can Remove Copyrighted Concepts from Text-to-Image Diffusion Models

Item Type	Blog Post
Author	Ekrem Çetinkaya
Abstract	Text-to-image models have stormed the AI domain in the last couple of months. They have demonstrated superb image generation performance, which can generate outputs using text prompts that can be difficult to distinguish from real images. These models are becoming an essential part of content generation quite quickly. Nowadays, it is possible to use AI models to generate images that we can use in our applications, let's say, webpage design. We can just take one of the models, which can be MidJourney, DALL-E, or Stable Diffusion, and ask them to generate images for us. Let us, for a second assume
Date	2023-04-06T16:13:23-07:00
Language	en-US
Short Title	Opt-Me-Out From Diffusion
URL	https://www.marktechpost.com/2023/04/06/opt-me-out-from-diffusion-this-ai-model-can-remove-copyrighted-concepts-from-text-to-image-diffusion-models/
Accessed	4/7/2023, 4:01:28 AM

Blog Title	MarkTechPost
Date Added	4/7/2023, 4:01:28 AM
Modified	4/7/2023, 4:01:35 AM

- ## Attachments

  - Snapshot

- # Optimizing Language Models for Argumentative Reasoning

Item Type	Journal Article
Author	Luke Thorburn
Author	Ariel Kruger
Date	2022
Library Catalog	Google Scholar
Extra	WFZ tldr: LLMs may be valuable part of a sensor fusion approach that includes other types of argument-enhancing technologies.
Date Added	9/15/2022, 5:53:10 AM
Modified	9/15/2022, 6:35:03 PM

- **Tags:**

  - ⊖ No DOI found

Attachments

  - Full Text

- # Paper page - Textbooks Are All You Need

Item Type	Web Page
URL	https://huggingface.co/papers/2306.11644
Accessed	6/21/2023, 5:44:15 AM
Date Added	6/21/2023, 5:44:15 AM
Modified	6/21/2023, 5:44:18 AM

- # Attachments

  - Paper page - Textbooks Are All You Need

- # Papercutting: An international bibliography and selected guide to US collections

Item Type	Book
Author	Martha Kreisel
Date	1994
Extra	WFZ Compare Midjourney-powered paperutting
Publisher	Scarecrow Press
Date Added	12/6/2022, 10:23:33 PM
Modified	12/15/2022, 8:54:59 AM

- ## Tags:

  - Books: history
  - Book Crafts
  - Papercutting

- # Passing Messages Between Disciplines

Item Type	Journal Article

- ## Tags:

- # [PDF] Enhancing Pre-Trained Language Representations Based on Contrastive Learning for Unsupervised Keyphrase Extraction | Semantic Scholar

Item Type	Web Page
URL	https://www.semanticscholar.org/paper/Enhancing-Pre-Trained-Language-Representation-Zhang/7f46c557fa1868823a2301bba43f4a9973df5c92?utm_source=alert_email&utm_co 0-1&utm_medium=7614946
Accessed	9/10/2022, 4:00:39 PM
Extra	WFZ Unsupervised keyword extraction is important to book discovery, indexing, and enh the keyphrase set should tend to be close to that of the extracted document, and far from th embeddings.
Date Added	9/10/2022, 4:00:39 PM
Modified	9/10/2022, 7:48:23 PM

- ## Attachments

  - ○ [PDF] Enhancing Pre-Trained Language Representations Based on Contrastive Learning for

- # [PDF] ISSUES IN MALAYALAM TEXT SUMMARIZATION | Semantic Scholar

Item Type	Web Page
URL	https://www.semanticscholar.org/paper/ISSUES-IN-MALAYALAM-TEXT-SUMMARIZ Mubarak/f35d0676c3c9cec1a6f10de1066fedafbf85cda2?utm_source=alert_email&utm_c 1-1&utm_medium=7595992
Accessed	9/9/2022, 2:27:09 PM
Date Added	9/9/2022, 2:27:09 PM
Modified	9/10/2022, 7:47:52 PM

- ## Attachments

  - [PDF] ISSUES IN MALAYALAM TEXT SUMMARIZATION | Semantic Scholar

- # [PDF] Learning to Paraphrase from Multi-Document Summarization Data Master's Thesis | Semantic Scholar

Item Type	Web Page
URL	https://www.semanticscholar.org/paper/Learning-to-Paraphrase-from-Multi-Document-Da Nguyen/d64ee98364f61da596e12232cd7a18a448d806bf?utm_source=alert_email&utm_c 3-5&utm_medium=7584502
Accessed	9/11/2022, 2:04:48 PM
Extra	WFZ Shows that fine-tuning on a paragraphs dataset improves paraphrasing by DistilBER
Date Added	9/11/2022, 2:04:48 PM
Modified	9/11/2022, 2:09:13 PM

- ## Attachments

  - [PDF] Learning to Paraphrase from Multi-Document Summarization Data Master's Thesis | Semantic Scholar

- # [PDF] Patterns of Book Ownership in Late Seventeenth-Century England | Semantic Scholar

Item Type	Journal Article
Abstract	A detailed comparative analysis of five late 17th-century English book auction catalogues was undertaken, by entering their contents (title by title) into a database. The collections were roughly similar in size but their owners had varying professional backgrounds, including establishment and nonconformist clergy, academic, and gentry. The 6900 lots represented ca.4200 individual

works by ca.2700 authors. By creating lists of titles and authors found in multiple catalogues, it was possible to draw inferences about the works which were commonly, and less commonly, held in private libraries of this period. The books most often found in these collections are not necessarily the ones which we value today as the most significant works of the early modern period. This approach helps us to gain a deeper and more accurate understanding of the contents of a typical scholarly library of this time.

Language	en
Library Catalog	www.semanticscholar.org
URL	https://www.semanticscholar.org/paper/Patterns-of-Book-Ownership-in-Late-England-Pearson/991772a34a076c8f220ad80dc00bae2f8fdd50de
Accessed	9/6/2022, 9:48:24 PM
Extra	WFZ "**The books most often found in these collections are not necessarily the ones which we value today** as the most significant works of the early modern period."
Date Added	9/6/2022, 9:48:24 PM
Modified	12/4/2022, 11:55:44 PM

- **Tags:**

    - #nosource

- # [PDF] Unsupervised Simplification of Legal Texts | Semantic Scholar

Item Type	Web Page
URL	https://www.semanticscholar.org/paper/Unsupervised-Simplification-of-Legal-Texts-Cem cCukur/e8bf54dc59a30da55828a028dbd738eebc325342?utm_source=alert_email&utm_c 4-6&utm_medium=7584502
Accessed	9/11/2022, 2:25:05 PM
Extra	WFZ Useful approach to smart tldr-ing. "USLT detects complex words in a sentence, gene score. Afterward, USLT recursively decomposes long sentences into a hierarchy of shorte state-of-the-art domain-general TS methods in text simplicity while keeping the semantics
Date Added	9/11/2022, 2:25:05 PM
Modified	9/11/2022, 3:03:27 PM

- **Attachments**

  - [PDF] Unsupervised Simplification of Legal Texts | Semantic Scholar

- # Penguin to publish "classic" Roald Dahl books after "censorship" backlash

Item Type	Web Page
Abstract	The book publisher had received criticism for removing passages related to weight, mental health, gender and race.

Date	2023-02-24
Language	en-US
URL	https://www.cbsnews.com/news/roald-dahl-penguin-original-books-censorship-backlash/
Accessed	2/26/2023, 12:25:34 AM
Date Added	2/26/2023, 12:25:34 AM
Modified	2/26/2023, 12:25:34 AM

- # **Attachments**

  - ○ Snapshot

- # Personalized Font Recommendations: Combining ML and Typographic Guidelines to Optimize Readability

Item Type	Conference Paper
Author	Tianyuan Cai
Author	Shaun Wallace
Author	Tina Rezvanian
Author	Jonathan Dobres

Author	Bernard Kerr
Author	Samuel Berlow
Author	Jeff Huang
Author	Ben D. Sawyer
Author	Zoya Bylinskii

Abstract The amount of text people need to read and understand grows daily. Software defaults, designers, or publishers often choose the fonts people read in. However, matching individuals with a faster font could help them cope with information overload. We collaborated with typographers to (1) select eight fonts designed for digital reading to systematically compare their effectiveness and to (2) understand how font and reader characteristics affect reading speed. We collected font preferences, reading speeds, and characteristics from 252 crowdsourced participants in a remote readability study. We use font and reader characteristics to train FontMART, a learning to rank model that automatically orders a set of eight fonts per participant by predicted reading speed. FontMART's fastest font prediction shows an average increase of 14–25 WPM compared to other font defaults, without hindering comprehension. This encouraging evidence provides motivation for adding our personalized font recommendation to future interactive systems.

Date	June 13, 2022
Short Title	Personalized Font Recommendations
Library Catalog	ACM Digital Library
URL	https://dl.acm.org/doi/10.1145/3532106.3533457
Accessed	6/26/2023, 8:00:00 PM
Place	New York, NY, USA

- **Tags:**

  - personalization
  - typography
  - readability
  - reading

Attachments

  - Full Text PDF

- # PGN-LM Model and Forcing-Seq2Seq Model: Multiple automatic models of

title generation for natural text using Deep Learning | Semantic Scholar

Item Type	Web Page
URL	https://www.semanticscholar.org/paper/PGN-LM-Model-and-Forcing-Seq2Seq-Model%3A-Multiple-of-Nhan-Thuan/044d8f842d4bc1a3fda6940342375ed88cff59c8
Accessed	9/6/2022, 9:45:25 PM
Date Added	9/6/2022, 9:45:25 PM
Modified	9/6/2022, 9:46:06 PM

- ## Attachments

  - PGN-LM Model and Forcing-Seq2Seq Model: Multiple automatic models of title generation for natural text using Deep Learning | Semantic Scholar

- # Poisoning Web-Scale Training Datasets is Practical

Item Type	Preprint
Author	Nicholas Carlini
Author	Matthew Jagielski

Author	Christopher A. Choquette-Choo
Author	Daniel Paleka
Author	Will Pearce
Author	Hyrum Anderson
Author	Andreas Terzis
Author	Kurt Thomas
Author	Florian Tramèr

Abstract Deep learning models are often trained on distributed, webscale datasets crawled from the internet. In this paper, we introduce two new dataset poisoning attacks that intentionally introduce malicious examples to a model's performance. Our attacks are immediately practical and could, today, poison 10 popular datasets. Our first attack, split-view poisoning, exploits the mutable nature of internet content to ensure a dataset annotator's initial view of the dataset differs from the view downloaded by subsequent clients. By exploiting specific invalid trust assumptions, we show how we could have poisoned 0.01% of the LAION-400M or COYO-700M datasets for just $60 USD. Our second attack, frontrunning poisoning, targets web-scale datasets that periodically snapshot crowd-sourced content -- such as Wikipedia -- where an attacker only needs a time-limited window to inject malicious examples. In light of both attacks, we notify the maintainers of each affected dataset and recommended several low-overhead defenses.

Date	2023-02-20
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2302.10149
Accessed	2/22/2023, 2:50:31 PM

Extra	arXiv:2302.10149 [cs]
DOI	10.48550/arXiv.2302.10149
Repository	arXiv
Archive ID	arXiv:2302.10149
Date Added	2/22/2023, 2:50:31 PM
Modified	2/22/2023, 2:50:33 PM

- **Tags:**

  - Computer Science - Machine Learning
  - Computer Science - Cryptography and Security

Attachments

  - arXiv Fulltext PDF
  - arXiv.org Snapshot

- # PolyFuzz: Fuzzy string matching, grouping, and evaluation.

Item Type	Document
Author	Maarten Grootendorst

Date	2020
URL	https://doi.org/10.5281/zenodo.4461050
Extra	Citation Key: grootendorst2020polyfuzz DOI: 10.5281/zenodo.4461050 tex.version: v0.2.2
Publisher	Zenodo
Date Added	12/23/2022, 5:02:16 PM
Modified	12/23/2022, 5:02:16 PM

- ## Prompt Reducer - Cut Down GPT-4 Token Costs

Item Type	Web Page
Abstract	Prompt Reducer helps you reduce GPT-4 prompts to a few lines, saving tokens and loweri
URL	https://promptreducer.com/
Accessed	4/10/2023, 9:38:03 AM
Extra	WFZ I include this because it is about language: GPT-4 can invent its own shorthand. best,worst.times,age.wisdom,foolishness.epoch.belief,incredulity.season.Light,Darkness.s
Date Added	4/10/2023, 9:38:03 AM
Modified	4/10/2023, 10:12:24 AM

- **Attachments**

  - Snapshot

promphero/openjourney – Run with an API on Replicate

Item Type	Web Page
URL	https://replicate.com/promphero/openjourney
Accessed	1/27/2023, 5:29:05 AM
Date Added	1/27/2023, 5:29:05 AM
Modified	1/27/2023, 5:29:05 AM

- **Attachments**

  - promphero/openjourney – Run with an API on Replicate

Prompting as Probing: Using Language Models for Knowledge Base Construction

Item Type	Preprint

Author	Dimitrios Alivanistos
Author	Selene Báez Santamaría
Author	Michael Cochez
Author	Jan-Christoph Kalo
Author	Emile van Krieken
Author	Thiviyan Thanapalasingam

Abstract Language Models (LMs) have proven to be useful in various downstream applications, such as summarisation, translation, question answering and text classification. LMs are becoming increasingly important tools in Artificial Intelligence, because of the vast quantity of information they can store. In this work, we present ProP (Prompting as Probing), which utilizes GPT-3, a large Language Model originally proposed by OpenAI in 2020, to perform the task of Knowledge Base Construction (KBC). ProP implements a multi-step approach that combines a variety of prompting techniques to achieve this. Our results show that manual prompt curation is essential, that the LM must be encouraged to give answer sets of variable lengths, in particular including empty answer sets, that true/false questions are a useful device to increase precision on suggestions generated by the LM, that the size of the LM is a crucial factor, and that a dictionary of entity aliases improves the LM score. Our evaluation study indicates that these proposed techniques can substantially enhance the quality of the final predictions: ProP won track 2 of the LM-KBC competition, outperforming the baseline by 36.4 percentage points. Our implementation is available on https://github.com/HEmile/iswc-challenge.

Date	2022-08-25
Language	en
Short Title	Prompting as Probing

Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2208.11057
Accessed	9/20/2022, 12:47:00 PM
Extra	WFZ Good discussion of prompting techniques that improve precision by 36% for fact identification use case. Formatting options as lists ['one', 'two', 'three'] helps, so does listing edge cases [].
Repository	arXiv
Archive ID	arXiv:2208.11057
Date Added	9/20/2022, 12:47:00 PM
Modified	9/20/2022, 12:48:51 PM

- **Tags:**

    ○ Computer Science - Computation and Language
    ○ Computer Science - Artificial Intelligence

Notes:

    ○ Comment: To be published in LM-KBC 22: Knowledge Base Construction from Pre-trained Language Models, Challenge at ISWC 2022. 12+12 pages

Attachments

PromptSource: An Integrated Development Environment and Repository for Natural Language Prompts

Item Type	Journal Article
Author	Stephen H. Bach
Author	Victor Sanh
Author	Zheng-Xin Yong
Author	Albert Webson
Author	Colin Raffel
Author	Nihal V. Nayak
Author	Abheesht Sharma
Author	Taewoon Kim
Author	M. Saiful Bari

Author   Thibault Fevry

Author   Zaid Alyafeai

Author   Manan Dey

Author   Andrea Santilli

Author   Zhiqing Sun

Author   Srulik Ben-David

Author   Canwen Xu

Author   Gunjan Chhablani

Author   Han Wang

Author   Jason Alan Fries

Author   Maged S. Al-shaibani

Author   Shanya Sharma

Author   Urmish Thakker

Author   Khalid Almubarak

Author   Xiangru Tang

Author   Xiangru Tang

Author   Mike Tian-Jian Jiang

Author	Alexander M. Rush

Abstract	PromptSource is a system for creating, sharing, and using natural language prompts. Prompts are functions that map an example from a dataset to a natural language input and target output. Using prompts to train and query language models is an emerging area in NLP that requires new tools that let users develop and refine these prompts collaboratively. PromptSource addresses the emergent challenges in this new setting with (1) a templating language for defining data-linked prompts, (2) an interface that lets users quickly iterate on prompt development by observing outputs of their prompts on many examples, and (3) a community-driven set of guidelines for contributing new prompts to a common pool. Over 2,000 prompts for roughly 170 datasets are already available in PromptSource. PromptSource is available at https://github.com/bigscience-workshop/promptsource.

Date	2022-02-02
Short Title	PromptSource
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2202.01279
Accessed	2/8/2022, 2:30:25 AM
Extra	arXiv: 2202.01279
Publication	arXiv:2202.01279 [cs]
Date Added	2/8/2022, 2:30:25 AM
Modified	12/4/2022, 11:55:10 PM

- # Tags:

- o #nosource
- o ⊖ No DOI found
- o Computer Science - Computation and Language
- o Computer Science - Machine Learning

- ## Provisions on the Administration of In-depth Synthesis of Internet Information Services

Item Type	Web Page
Author	Office of the Cyber Security and Informatization Commission of the Central Committee of the Communist Party of China
Date	2022-12-11
URL	http://www.cac.gov.cn/2022-12/11/c_1672221949354811.htm
Accessed	12/13/2022, 3:25:22 AM
Extra	WFZ Full text in Chinese with machine translation available.
Date Added	12/13/2022, 3:25:22 AM
Modified	12/15/2022, 8:45:57 AM

- ## Tags:

- o China

- ○ law \& legislation
  - ○ Safety

Attachments

  - ○ Provisions on the Administration of In-depth Synthesis of Internet Information Services

---

• PUBLISHER OPERATING MANUAL

Item Type	Document
Author	Lightning Source, Inc.
Date	October 26, 2022
URL	https://myaccount.lightningsource.com/documents/LSI/files/pod/PODOpsManual.pdf
Accessed	3/5/2023, 3:10:50 AM
Extra	WFZ Version 4.6.
Publisher	Lightning Source, Inc.
Date Added	3/5/2023, 3:12:04 AM
Modified	3/5/2023, 3:13:29 AM

• **Attachments**

- PODOpsManual.pdf

Publishers want billions, not millions, from AI | Semafor

Item Type	Web Page
Author	https://www.semafor.com/author/ben-smith
Abstract	Barry Diller fired publishers' opening shot at artificial intelligence platforms in a Semafor interview this April, suggesting they sue the companies that have trained models on their data.
Date	2023-07-24T01:03:05Z
Language	en
URL	https://www.semafor.com/article/07/23/2023/publishers-want-billions-not-millions-from-ai
Accessed	7/24/2023, 3:46:10 AM
Extra	WFZ Publishers are looking for billions from tech companies this time around, that's "B" as in billion.
Date Added	7/24/2023, 3:46:10 AM
Modified	7/27/2023, 2:24:22 AM

- # Publishers' group warns that generative AI content could violate copyright law

Item Type	Web Page
Abstract	A group whose members include the New York Times and the Washington Post has advice on how to navigate concerns surrounding AI.
Language	en-us
URL	https://www.marketingbrew.com/stories/2023/06/05/publishers-group-warns-that-generative-ai-content-could-violate-copyright-law
Accessed	6/6/2023, 12:18:40 AM
Website Title	Marketing Brew
Date Added	6/6/2023, 12:18:39 AM
Modified	6/6/2023, 12:18:39 AM

- # **Attachments**

  - Snapshot

- # Quantitative Analysis of Culture Using Millions of Digitized Books

Item Type	Journal Article
Author	Jean-Baptiste Michel
Author	Yuan Shen
Author	Aviva Aiden
Author	Adrian Veres
Author	Matthew Gray
Author	Joseph Pickett
Author	Dale Hoiberg
Author	Dan Clancy
Author	Peter Norvig
Author	Jon Orwant
Author	Steven Pinker
Author	Martin Nowak
Author	Erez Aiden

Abstract We constructed a corpus of digitized texts containing about 4% of all books ever printed. Analysis of this corpus enables us to investigate cultural trends quantitatively. We survey the vast terrain of 'culturomics,' focusing on linguistic and cultural phenomena that were reflected in the English language between 1800 and 2000. We show how this approach can provide insights about fields as diverse as lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical

epidemiology. Culturomics extends the boundaries of rigorous quantitative inquiry to a wide array of new phenomena spanning the social sciences and the humanities.

Date	2011-01-14
Library Catalog	ResearchGate
Volume	331
Pages	176-82
Publication	Science (New York, N.Y.)
DOI	10.1126/science.1199644
Journal Abbr	Science (New York, N.Y.)
Date Added	6/7/2022, 11:42:43 AM
Modified	6/18/2022, 6:44:48 PM

- **Attachments**

  - Full Text PDF
  - ResearchGate Link

- # Quantitative analysis of fanfictions' popularity

- **Tags:**

  - #nosource

- # Query-Focused Re-Ranking to Enhance the Performance of Text Entailment and Question Answering

Item Type	Conference Paper
Author	Vikanksh Nath
Author	Vipul Chauhan
Author	Lipika Dey
Abstract	Transformer-based models have dramatically improved performance of various natural language processing tasks like question answering, fact verification, topic-driven summarization and natural language inferencing. However, these models can't process input context longer than their token-length limit (TLL) at a time. Given a large document however, the required context may be spread over a larger area and also may not be restricted to contiguous sentences. Existing methods fail to handle such situations correctly. In this paper, we propose a method to handle this issue by detecting the right context from a large document before performing the actual query-context text-pair task. The proposed method fragments a long text document into sub-texts and then employs a cross-encoder model to generate a query-

focused relevance score for each sub-text module. The actual downstream task is performed with the most relevant sub-text chosen as the context, rather than arbitrarily selecting the top few sentences. This extricates the model from the traditional way of iterating over TLL window size text fragments and saves computational cost. The efficacy of the approach has been established with multiple tasks. The proposed model out-performs several state of the art models for the tasks by a significant margin.

Date   January 4, 2023

Library Catalog   ACM Digital Library

URL   https://doi.org/10.1145/3570991.3571015

Accessed   2/17/2023, 7:00:00 PM

Place   New York, NY, USA

Publisher   Association for Computing Machinery

ISBN   978-1-4503-9797-1

Pages   47–54

Series   CODS-COMAD '23

Proceedings Title   Proceedings of the 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD)

DOI   10.1145/3570991.3571015

Date Added   2/18/2023, 6:35:29 PM

Modified   2/18/2023, 6:35:29 PM

- **Tags:**

  - language model
  - long text
  - natural language inference
  - question answering
  - re-ranking
  - transformers

- # Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data

Item Type	Conference Paper
Author	Xintao Wang
Author	Liangbin Xie
Author	Chao Dong
Author	Ying Shan
Date	2021
Extra	WFZ Very good choice for upscaling images from Stable Diffusion or DALLE to size needed for book covers. Available as standalone app for Win/Lin/OSX, also as Python library and cli.
Proceedings Title	International conference on computer vision workshops (ICCVW)

- # Redefining what a map can be with new information and AI

Item Type	Web Page
Author	Dane Glasgow
URL	https://blog.google/products/maps/redefining-what-map-can-be-new-information-and-ai
Accessed	3/30/2021, 8:00:00 PM
Website Title	Google.
Date Added	3/31/2021, 6:09:45 PM
Modified	12/4/2022, 11:55:06 PM

- ## Tags:

  - #nosource

- # Refining Character Relationships in a Narrative using Embeddings of Interactions

Item Type	Journal Article
Author	Guillaume Guex
Abstract	The construction of character networks from narratives can be decomposed into three steps: (1) the identification of characters; (2) the detection of interactions; (3) the extraction of the graph (for an extensive review, see (Labatut and Bost, 2019)). When the studied narrative consists in textual resources (e.g. novels, plays, movie scripts), the detection of interactions (step 2) between two characters generally consists in identifying segments of text where protagonists are involved in some type of interaction. These segments of text can be, among others, narrative units containing character co-occurrences (Elsner, 2012; Rochat and Kaplan, 2014; Min and Park, 2019), conversations (Nalisnick and Baird, 2013; Kwon and Shim, 2017), or direct actions (Sudhahar and Cristianini, 2013; Srivastava et al., 2016). In step 3, these interactions are usually counted to obtain an undirected/directed, static/dynamic weighted character network where edges represent the number of interactions. However, apart from a few examples where edges are signed according to a sentiment score (Nalisnick and Baird, 2013; Trovati and Brady, 2014; Min and Park, 2019), almost no information is extracted from the corpus formed by the segments of text defining interactions. Interpreting relationships in a resulting network can be therefore difficult, as edges aggregate blindly various types of interactions.
Language	en
Library Catalog	Zotero
Pages	3
Date Added	6/22/2022, 5:51:59 PM

Modified	6/22/2022, 5:52:00 PM

- **Tags:**

  - ⊖ No DOI found

Attachments

  - Guex - Refining Character Relationships in a Narrative usi.pdf

---

- ## "Regulations on the Administration of Deep Synthesis of Internet Information Services" Answers to Reporters' Questions-

Item Type	Web Page
Author	Office of the Cyber Security and Informatization Commission of the Central Committee of the Communist Party of China
Date	2022-12-11
URL	http://www.cac.gov.cn/2022-12/11/c_1672221949570926.htm
Accessed	12/13/2022, 3:25:46 AM
Extra	WFZ FAQ from PRC officials.

Date Added	12/13/2022, 3:25:46 AM
Modified	12/15/2022, 8:45:32 AM

- ## Tags:

  - China
  - Safety

Attachments

- "Regulations on the Administration of Deep Synthesis of Internet Information Services" Answers to Reporters' Questions-Office of the Cyber Security and Informatization Commission of the Central Committee of the Communist Party of China

---

- # Representing Movie Characters in Dialogues

Item Type	Conference Paper
Author	Mahmoud Azab
Author	Noriyuki Kojima
Author	Jia Deng

Author	Rada Mihalcea
Abstract	We introduce a new embedding model to represent movie characters and their interactions in a dialogue by encoding in the same representation the language used by these characters as well as information about the other participants in the dialogue. We evaluate the performance of these new character embeddings on two tasks: (1) character relatedness, using a dataset we introduce consisting of a dense character interaction matrix for 4,761 unique character pairs over 22 hours of dialogue from eighteen movies; and (2) character relation classification, for fine- and coarse-grained relations, as well as sentiment relations. Our experiments show that our model significantly outperforms the traditional Word2Vec continuous bag-of-words and skip-gram models, demonstrating the effectiveness of the character embeddings we introduce. We further show how these embeddings can be used in conjunction with a visual question answering system to improve over previous results.
Date	2019
Language	en
Library Catalog	DOI.org (Crossref)
URL	https://www.aclweb.org/anthology/K19-1010
Accessed	8/22/2022, 10:33:26 PM
Place	Hong Kong, China
Publisher	Association for Computational Linguistics
Pages	99-109
Proceedings Title	Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)

Conference Name	Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)
DOI	
Date Added	8/22/2022, 10:33:26 PM
Modified	8/22/2022, 10:33:27 PM

- **Attachments**

  - Azab et al. - 2019 - Representing Movie Characters in Dialogues.pdf

- # Retentive Network: A Successor to Transformer for Large Language Models

Item Type	Preprint
Author	Yutao Sun
Author	Li Dong
Author	Shaohan Huang
Author	Shuming Ma
Author	Yuqing Xia

Author	Jilong Xue
Author	Jianyong Wang
Author	Furu Wei

Abstract In this work, we propose Retentive Network (RetNet) as a foundation architecture for large language models, simultaneously achieving training parallelism, low-cost inference, and good performance. We theoretically derive the connection between recurrence and attention. Then we propose the retention mechanism for sequence modeling, which supports three computation paradigms, i.e., parallel, recurrent, and chunkwise recurrent. Specifically, the parallel representation allows for training parallelism. The recurrent representation enables low-cost $O(1)$ inference, which improves decoding throughput, latency, and GPU memory without sacrificing performance. The chunkwise recurrent representation facilitates efficient long-sequence modeling with linear complexity, where each chunk is encoded parallelly while recurrently summarizing the chunks. Experimental results on language modeling show that RetNet achieves favorable scaling results, parallel training, low-cost deployment, and efficient inference. The intriguing properties make RetNet a strong successor to Transformer for large language models. Code will be available at https://aka.ms/retnet.

Date	2023-07-19
Short Title	Retentive Network
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2307.08621
Accessed	7/20/2023, 6:56:20 PM
Extra	WFZ Old: LLM. Bold: RTN. Many levels of performance improvement => better at processing large texts.

- **Tags:**

    - Computer Science - Computation and Language
    - Computer Science - Machine Learning

Attachments

    - arXiv Fulltext PDF
    - arXiv.org Snapshot

- # Retrieval for Extremely Long Queries and Documents with RPRS: a Highly Efficient and Effective Transformer-based Re-Ranker

Item Type	Conference Paper
Author	Arian Askari

Author	S. Verberne
Author	Amin Abolghasemi
Author	Wessel Kraaij
Author	G. Pasi
Abstract	Retrieval with extremely long queries and documents is a well-known and challengin designed Transformer models that can handle long input sequences have not shown h Relevance Score (RPRS) to compute the relevance score between a query and the top state-of-the-art models on five different datasets. Furthermore, RPRS is highly efficie ranker the advantage of having a complexity of O(N) where N is the total number of s resource training in QBD retrieval tasks as it does not need large amounts of training amount of labeled data is available. Our detailed analysis shows that RPRS benefits f
Date	2 March 2023
Short Title	Retrieval for Extremely Long Queries and Documents with RPRS
Library Catalog	Semantic Scholar
URL	https://www.semanticscholar.org/paper/Retrieval-for-Extremely-Long-Queries-and-D Verberne/55fc37551364e27dec71f363da9b99e950667dd2?utm_source=alert_email& 0-2&utm_medium=12787004
Accessed	3/5/2023, 9:46:22 AM
Date Added	3/5/2023, 9:46:22 AM
Modified	3/5/2023, 9:46:22 AM

- ## **Notes:**

- o [TLDR] This work proposes a Re-Ranker based on the novel Proportional Relevance Score (RPRS) to compute the relevance score between a query and the top-k candidate documents, and shows that RPRS benefits from covering the full length of candidate documents and queries.

Attachments

- o Full Text PDF
- o Semantic Scholar Link

Retrieval-efficiency trade-off of unsupervised keyword extraction

Item Type	Journal Article
Author	Blaž Škrlj
Author	Boshko Koloski
Author	Senja Pollak
Date	2022
Extra	WFZ Claims to be 100x faster than YAKE, which I currently use.
Volume	abs/2208.07262
Publication	ArXiv

Date Added	8/24/2022, 3:46:53 PM
Modified	12/4/2022, 11:55:43 PM

- **Tags:**

  - #nosource
  - ⊖ No DOI found

- # Revisiting Indexing & Abstracting in the Digital Era

Item Type	Journal Article
Author	Samantha King
Author	Howard Boyedoe
Author	Andrea Chacon
Author	Molli Hall
Author	Maria Krizansky
Author	Kathleen O'Bryant
Author	D. Alemneh
Date	2018

Library Catalog	Google Scholar
Date Added	8/11/2022, 7:33:35 AM
Modified	8/11/2022, 7:33:40 AM

- **Tags:**

  - ⊖ No DOI found

Attachments

  - Full Text

- # RL with KL penalties is better viewed as Bayesian inference

Item Type	Preprint
Author	Tomasz Korbak
Author	Ethan Perez
Author	Christopher L. Buckley
Abstract	Reinforcement learning (RL) is frequently employed in fine-tuning large language models (LMs) to penalize them for undesirable features of generated sequences, such as offensiveness or harmfulness. In this paper, we analyze challenges associated with treating language models as RL policies and show how avoiding those challenges requires moving beyond the RL paradigm.

- ## Tags:

  - Computer Science - Machine Learning
  - Statistics - Machine Learning

Notes:

- Comment: Findings of EMNLP 2022

Attachments

- Korbak et al. - 2022 - RL with KL penalties is better viewed as Bayesian .pdf

Robo-writers: the rise and risks of language-generating AI

Item Type	Journal Article
Author	Matthew Hutson
Date	2021-03-04
URL	http://www.nature.com/articles/d41586-021-00530-0
Accessed	3/2/2021, 7:00:00 PM
Volume	591
Pages	22-25
Publication	Nature
DOI	10.1038/d41586-021-00530-0

Issue	7848
ISSN	0028-0836
Date Added	3/3/2021, 8:03:30 PM
Modified	12/4/2022, 11:50:38 PM

- **Tags:**

  - #nosource

S2ORC: The Semantic Scholar Open Research Corpus

Item Type	Journal Article
Author	Kyle Lo
Author	Lucy Lu Wang
Author	Mark Neumann
Author	Rodney Kinney
Author	Daniel Weld
Abstract	We introduce S2ORC, a large corpus of 81.1M English-language academic papers spanning many academic disciplines. The corpus consists of rich metadata, paper abstracts, resolved bibliographic references, as well as structured full text for 8.1M open access papers. Full text is annotated with automatically-detected inline

mentions of citations, figures, and tables, each linked to their corresponding paper objects. In S2ORC, we aggregate papers from hundreds of academic publishers and digital archives into a unified source, and create the largest publicly-available collection of machine-readable academic text to date. We hope this resource will facilitate research and development of tools and tasks for text mining over academic text.

Date	2020-07-29
Accessed	10/24/2021, 8:00:00 PM
Extra	WFZ Semantic Scholar is an indispensable resource combining very high quality scientific content with mature, stable, and high-quality data mining tools. Includes some very useful specialized tools for PDF manipulation and citation extraction.
Pages	4969-4983
DOI	10.18653/V1/2020.ACL-MAIN.447
Date Added	10/25/2021, 11:07:51 PM
Modified	12/4/2022, 11:55:08 PM

- **Tags:**

  - #nosource

- # Scaling Autoregressive Models for Content-Rich Text-to-Image Generation

Item Type Preprint

Author Jiahui Yu

Author Yuanzhong Xu

Author Jing Yu Koh

Author Thang Luong

Author Gunjan Baid

Author Zirui Wang

Author Vijay Vasudevan

Author Alexander Ku

Author Yinfei Yang

Author Burcu Karagol Ayan

Author Ben Hutchinson

Author Wei Han

Author Zarana Parekh

Author Xin Li

Author Han Zhang

Author Jason Baldridge

Author	Yonghui Wu
Abstract	We present the Pathways Autoregressive Text-to-Image (Parti) model, which generates high-fidelity photorealistic images and supports content-rich synthesis involving complex compositions and world knowledge. Parti treats text-to-image generation as a sequence-to-sequence modeling problem, akin to machine translation, with sequences of image tokens as the target outputs rather than text tokens in another language. This strategy can naturally tap into the rich body of prior work on large language models, which have seen continued advances in capabilities and performance through scaling data and model sizes. Our approach is simple: First, Parti uses a Transformer-based image tokenizer, ViT-VQGAN, to encode images as sequences of discrete tokens. Second, we achieve consistent quality improvements by scaling the encoder-decoder Transformer model up to 20B parameters, with a new state-of-the-art zero-shot FID score of 7.23 and finetuned FID score of 3.22 on MS-COCO. Our detailed analysis on Localized Narratives as well as PartiPrompts (P2), a new holistic benchmark of over 1600 English prompts, demonstrate the effectiveness of Parti across a wide variety of categories and difficulty aspects. We also explore and highlight limitations of our models in order to define and exemplify key areas of focus for further improvements. See https://parti.research.google/ for high-resolution images.
Date	2022-06-21
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2206.10789
Accessed	6/30/2022, 11:15:05 PM
Extra	arXiv:2206.10789 [cs]
DOI	10.48550/arXiv.2206.10789
Repository	arXiv
Archive ID	arXiv:2206.10789

Date Added	6/30/2022, 11:15:05 PM
Modified	6/30/2022, 11:15:05 PM

- ## Tags:

  - Computer Science - Machine Learning
  - Computer Science - Computer Vision and Pattern Recognition

Notes:

  - Comment: Preprint

Attachments

  - arXiv Fulltext PDF
  - arXiv.org Snapshot

- # Sci-Fi Mag Pauses Submissions Amid Flood of AI-Generated Short Stories

Item Type	Web Page
Abstract	Clarkesworld speculates that people are using AI to make 'a quick buck' since the sci-fi publication pays 12 cents a word for accepted stories.
Language	en

URL	https://www.pcmag.com/news/sci-fi-mag-pauses-submissions-amid-flood-of-ai-generated-short-stories
Accessed	2/22/2023, 9:03:52 AM
Extra	WFZ AI = death to over-the-transom submissions => more reliance on informal networks => more power to agents and publishers.
Website Title	PCMAG
Date Added	2/22/2023, 9:03:52 AM
Modified	2/22/2023, 9:27:47 AM

- **Attachments**

  - Snapshot

- # Scientists are using AI to dream up revolutionary new proteins

Item Type	Journal Article
Abstract	Huge advances in artificial intelligence mean researchers can design completely original molecules in seconds instead of months.
Date	2022-09-15
Language	en

- **Tags:**

  - Proteomics
  - Structural biology

Attachments

  - Snapshot

- # Segment Anything

Item Type	Preprint
Author	Alexander Kirillov
Author	Eric Mintun
Author	Nikhila Ravi
Author	Hanzi Mao
Author	Chloe Rolland
Author	Laura Gustafson
Author	Tete Xiao
Author	Spencer Whitehead
Author	Alexander C. Berg
Author	Wan-Yen Lo
Author	Piotr Dollár
Author	Ross Girshick

Abstract   We introduce the Segment Anything (SA) project: a new task, model, and dataset for image segmentation. Using our efficient model in a data collection loop, we built the largest segmentation dataset to date (by far), with over 1 billion masks on 11M licensed and privacy respecting images. The model is designed and trained to be promptable, so it can transfer zero-shot to new image distributions and tasks. We evaluate its capabilities on numerous tasks and find that its zero-shot performance is impressive -- often competitive with or even superior to prior fully supervised results. We are releasing the Segment Anything Model (SAM) and corresponding dataset (SA-1B) of 1B masks and

11M images at https://segment-anything.com to foster research into foundation models for computer vision.

Date	2023-04-05
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2304.02643
Accessed	4/12/2023, 2:30:47 PM
Extra	arXiv:2304.02643 [cs]
DOI	10.48550/arXiv.2304.02643
Repository	arXiv
Archive ID	arXiv:2304.02643
Date Added	4/12/2023, 2:30:47 PM
Modified	4/12/2023, 2:30:47 PM

- ## Tags:

  - Computer Science - Machine Learning
  - Computer Science - Artificial Intelligence
  - Computer Science - Computer Vision and Pattern Recognition

Notes:

- Comment: Project web-page: https://segment-anything.com

Attachments

- arXiv Fulltext PDF
- arXiv.org Snapshot

Self-Instruct: Aligning LM with Self Generated Instructions

Item Type	Software
Programmer	Yizhong Wang
Abstract	Aligning pretrained language models with instruction data generated by themselves.
Date	2023-03-15T21:31:43Z
Short Title	Self-Instruct
Library Catalog	GitHub
URL	https://github.com/yizhongw/self-instruct
Accessed	3/15/2023, 5:43:38 PM

Rights	Apache-2.0
Extra	original-date: 2022-12-20T08:52:51Z
Prog. Language	Python
Date Added	3/15/2023, 5:43:38 PM
Modified	3/15/2023, 5:43:43 PM

- **Tags:**

  - general-purpose-model
  - instruction-tuning
  - language-model

- # Semantic Overlap Summarization among Multiple Alternative Narratives: An Exploratory Study

Item Type	Journal Article
Author	Naman Bansal
Author	Mousumi Akter
Author	Shubhra Kanti Karmaker
Language	en

Library Catalog	Zotero
Pages	13
Date Added	10/16/2022, 7:50:52 AM
Modified	10/16/2022, 7:50:53 AM

- **Tags:**

  - ⊖ No DOI found

Attachments

  - Bansal et al. - Semantic Overlap Summarization among Multiple Alte.pdf

- # Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text

Item Type	Journal Article
Author	Saif M. Mohammad
Date	2021-01-01
Accessed	9/2/2021, 8:00:00 PM

Extra	Publisher: Woodhead Publishing
Pages	323-379
Publication	Emotion Measurement
DOI	[10.1016/B978-0-12-821124-3.00011-9](10.1016/B978-0-12-821124-3.00011-9)
Date Added	9/3/2021, 7:53:26 AM
Modified	12/4/2022, 11:55:07 PM

- ## Tags:

  - #nosource

- # Setting Opinion Aside - An AI Model That Ranks Speeches by Interest —

Item Type	Web Page
Abstract	Treegoat Media set out to determine if a deep learning artificial intelligence (AI) model could provide a statistical method of scoring how "interesting" a speech is based solely on the text. Speeches represent an important communication format for a wide range of government, business, and social ac
Language	en-US
URL	https://www.treegoatmedia.com/whitepaper-setting-opinion-aside
Accessed	11/17/2022, 3:03:53 PM

Extra	WFZ Treegoat and Voicebot.AI trained a model on ~120 longform public speeches of various types and devised its own proprietary metric for "interestingness" to apply to the text of speeches, passages, and sentences. No human coding or source of truth, so the results are debatable, but interesting because they are generated by the same tests for each document. Their metric finds that increased subjectivity is highly correlated with interestingness, and corresponds with a shift towards increased subjectivity in public speech over the last few decades. (Which in turn is a probably a reflection of broader social trends whose value is also arguable!) White paper available on request. This work could easily be extended to longform text in books. I have my own ideas about how to measure interestingness, but this is a good effort. Their first product offering will be "Marbyls", or interesting passages extracted from podcasts; due out later this year.
Website Title	Treegoat
Date Added	11/17/2022, 3:03:53 PM
Modified	11/17/2022, 4:12:15 PM

- ## Attachments

  - Setting Opinion Aside - An AI Model That Ranks Spe.pdf
  - Snapshot

- # Shutterstock to integrate OpenAI's DALL-E 2 and launch fund for contributor artists | TechCrunch

Item Type	Web Page

URL	https://techcrunch.com/2022/10/25/shutterstock-openai-dall-e-2/
Accessed	10/28/2022, 11:46:26 PM
Extra	WFZ A book publisher will do this soon. Maybe a genre publisher like Harlequin or Tor. I'm already doing it at Nimble, just behind the scenes.
Date Added	10/28/2022, 11:46:26 PM
Modified	10/29/2022, 7:11:55 PM

- ## Attachments

  - ○ Shutterstock to integrate OpenAI's DALL-E 2 and launch fund for contributor artists | TechCrunch

- # Simon Pegg says AI will weed out mediocrity in Hollywood

Item Type	Web Page
Abstract	Its use has prompted the biggest strike by Hollywood actors in decades, with A-list stars joining writers and extras on studio picket lines. But the use of Artificial intelligence in film and TV production has found qualified support from an unlikely quarter – Mission Impossible star Simon Pegg. The British actor and writer has said that the threat...
Date	1689441020
Language	en-gb

URL	https://www.msn.com/en-gb/entertainment/movies/simon-pegg-says-ai-as-a-tool-for-filmmaking-will-weed-out-mediocrity-in-hollywood/ar-AA1dUv3Z
Accessed	7/17/2023, 9:58:36 AM
Extra	WFZ Simon Pegg sticking his neck out by stating the obvious.
Date Added	7/17/2023, 9:58:36 AM
Modified	7/21/2023, 9:38:28 PM

- ## Notes:

  - But Pegg, who stars as Benji Dunn alongside Tom Cruise in the new Mission Impossible: Dead Reckoning, said that, when it came to the cinematic creative process itself, AI lacks the emotion, empathy and imagination required to create good art. "I think the writing process, it is a process, and when you write a first draft you write something that you know is going to improve and you will improve," said the 53-year-old. "If we get AI to write those first drafts the whole time people are only ever going to be doctoring scripts or giving notes.

Attachments

  - Snapshot

Social network analysis of Japanese manga: similarities to real-world social networks and trends over decades

Item Type	Preprint
Author	Kashin Sugishita
Author	Naoki Masuda
Abstract	Manga, Japanese comics, has been popular on a global scale. Social networks among characters, which are often called character networks, may be a significant contributor to their popularity. We collected data from 162 popular manga that span over 70 years and analyzed their character networks. First, we found that many of static and temporal properties of the character networks are similar to those of real human social networks. Second, the character networks of most manga are protagonist-centered such that a single protagonist interacts with the majority of other characters. Third, the character networks for manga mainly targeting boys have shifted to denser and less protagonist-centered networks and with fewer characters over decades. Manga mainly targeting girls showed the opposite trend except for the downward trend in the number of characters. The present study, which relies on manga data sampled on an unprecedented scale, paves the way for further population studies of character networks and other aspects of comics.
Date	2023-03-13
Short Title	Social network analysis of Japanese manga
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2303.07208

Accessed	3/15/2023, 9:51:59 PM
Extra	arXiv:2303.07208 [physics]
Repository	arXiv
Archive ID	arXiv:2303.07208
Date Added	3/15/2023, 9:51:59 PM
Modified	3/15/2023, 9:51:59 PM

- **Tags:**

  - Computer Science - Social and Information Networks
  - Physics - Physics and Society

Notes:

  - Comment: 9 pages, 7 figures

Attachments

  - arXiv Fulltext PDF
  - arXiv.org Snapshot

- # Soon You'll Be Able to Make Your Own Movie With AI Artificial

intelligence isn't about to change the movie industry. It already has.

Item Type	Magazine Article
Author	Lane Brown
Date	2022-12-27
URL	https://www.vulture.com/2022/12/ai-art-midjourney-chatgpt-phenaki-movies-hollywood.html
Extra	WFZ Movies first, then books.
Publication	New York
Date Added	12/30/2022, 8:28:48 AM
Modified	12/30/2022, 8:28:48 AM

- ## Stability AI Announces $101 Million in Funding for Open-Source Artificial Intelligence

Item Type	Web Page
Author	Stability.AI
Abstract	/PRNewswire/ -- Stability AI , the world's first community-driven, open-source artificial intelligence (AI) company today announced USD $101 million in funding....

Language	en
URL	https://www.prnewswire.com/news-releases/stability-ai-announces-101-million-in-funding-for-open-source-artificial-intelligence-301650932.html
Accessed	10/17/2022, 4:57:11 PM
Extra	WFZ Stability AI's release of Stable Diffusion was the best thing to happen in the AI world all year. Corporate gatekeepers are death to creativity. The news that there is $101M more to do stuff like that is great for book-lovers.
Date Added	10/17/2022, 4:57:11 PM
Modified	10/17/2022, 11:26:37 PM

- ## **Attachments**

  - Snapshot

- # Stable Diffusion launch announcement

Item Type	Web Page
Author	Stability.AI
Abstract	Stable Diffusion launch announcement
Language	en-GB
URL	https://stability.ai/blog/stable-diffusion-announcement

Accessed	8/13/2022, 5:45:09 AM
Website Title	Stability.Ai
Date Added	8/13/2022, 5:45:09 AM
Modified	8/13/2022, 10:38:49 PM

- # **Attachments**

  - ○ Snapshot

- # Stanford alpaca: An instruction-following LLaMA model

Item Type	Document
Author	Rohan Taori
Author	Ishaan Gulrajani
Author	Tianyi Zhang
Author	Yann Dubois
Author	Xuechen Li
Author	Carlos Guestrin

Author	Percy Liang
Author	Tatsunori B. Hashimoto
Date	2023
URL	https://github.com/tatsu-lab/stanford_alpaca
Extra	Citation Key: alpaca
Publisher	GitHub
Date Added	5/7/2023, 12:59:36 AM
Modified	5/7/2023, 12:59:36 AM

- # Startups and the technique behind ChatGPT

Item Type	Web Page
Author	Fraser
Abstract	Entropy #2
Date	2023-01-26
Language	en
URL	https://www.moreentropy.com/p/startups-and-the-technique-behind?publication_id=1338585&utm_medium=email%2Cemail&comments=true

Accessed	1/26/2023, 2:12:05 PM
Date Added	1/26/2023, 2:12:04 PM
Modified	1/26/2023, 2:12:04 PM

- **Attachments**

  - Snapshot

- # State of the Art in Algorithmic Publishing, 2023 ("SOTAPUB"): Books Created In Silico Using AI, Machine Learning, and Other Software Tools

Item Type	Book
Author	Frederick Zimmerman
Date	January 10, 2023
Language	English
Short Title	SOTAPUB 2023
URL	https://NimbleBooks.com

Extra	WFZ Software, presented as a book, about how to make books, using software. _In progress_, will publish in 2023.
Place	Ann Arbor
Publisher	Nimble Books LLC
ISBN	978-1-60888-199-4
Date Added	8/2/2022, 5:34:01 AM
Modified	12/4/2022, 11:55:38 PM

- ## Tags:
  - ### #nosource

- # Supporting content decision makers with machine learning

Item Type	Web Page
Author	Netflix Technology Blog
Abstract	by Melody Dye, Chaitanya (Chaitu) Ekanadham, Avneesh Saluja, and Ashish Rastogi
Date	2021-05-10T19:55:10.530Z
Language	en

URL	https://netflixtechblog.com/supporting-content-decision-makers-with-machine-learning-995b7b76006f
Accessed	9/25/2022, 1:32:43 AM
Extra	WFZ Standard ML techniques using large database of historical "content" performance. Key questions they try to answer: Which existing titles are comparable and in what ways? What audience size can we expect and in which regions? Embeddings used to create various types of similarity maps. https://miro.medium.com/max/1400/1*f2aTbsLsMaSFa0JcTffJfg.png Relies heavily on retrospective comparisons to existing titles
Website Title	Medium
Date Added	9/25/2022, 1:32:43 AM
Modified	9/25/2022, 1:43:21 AM

- # **Attachments**

  - Snapshot

- # Supporting diverse books through the Thema classification system - BookMachine

Item Type	Web Page
Abstract	At Edelweiss, we've seen that behind every successful book stands good metadata. Accurate and comprehensive information about a title's contents and themes is

Date	2022-10-17T08:00:00+01:00
Language	en-US
URL	https://bookmachine.org/2022/10/17/supporting-diverse-books-through-the-thema-classification-system/
Accessed	10/25/2022, 4:07:36 AM
Extra	WFZ Good discussion of how Thema can help book readers and sellers create very specific niche targets -- maybe even too specific!
Date Added	10/25/2022, 4:07:36 AM
Modified	10/26/2022, 2:57:21 PM

- # Attachments

  - ○ Snapshot

- # Survey of hallucination in natural language generation

Item Type	Document
Author	Ziwei Ji
Author	Nayeon Lee
Author	Rita Frieske

Author	Tiezheng Yu
Author	Dan Su
Author	Yan Xu
Author	Etsuko Ishii
Author	Yejin Bang
Author	Andrea Madotto
Author	Pascale Fung

Date	2022
URL	https://arxiv.org/abs/2202.03629
Extra	WFZ "Hallucination" of made-up and often impossible facts is a major problem for long-form generations since even a single ridiculously implausible statement can destroy the credibility of an entire work. But I don't like the connotations of "hallucination". One of the major strengths of transformers, in my view, is the ability to go beyond the confines not just of reality, but of the human imagination.
Publisher	arXiv
Date Added	8/5/2022, 9:51:01 PM
Modified	12/4/2022, 11:55:38 PM

- # Tags:

- ## SynthBio: A Case Study in Human-AI Collaborative Curation of Text Datasets

Item Type	Journal Article
Author	Ann Yuan
Author	Google Research
Author	Daphne Ippolito
Author	Vitaly Nikolaev
Author	Chris Callison-Burch
Author	Andy Coenen
Author	Sebastian Gehrmann
Abstract	NLP researchers need more, higher-quality text datasets. Human-labeled datasets are expensive to collect, while datasets collected via automatic retrieval from the web such as WikiBio [32] are noisy and can include undesired biases. Moreover, data sourced from the web is often included in datasets used to pretrain models, leading to inadvertent cross-contamination of training and test sets. In this work we introduce a novel method for efficient dataset curation: we use a large language model to provide seed generations to human raters, thereby changing dataset authoring from a writing task to an editing task. We use our method to curate

SynthBio-a new evaluation set for WikiBio-composed of structured attribute lists describing fictional individuals, mapped to natural language biographies. We show that our dataset of fictional biographies is less noisy than WikiBio, and also more balanced with respect to gender and nationality.

Accessed	12/31/2021, 7:00:00 PM
Extra	arXiv: 2111.06467v1
Date Added	1/1/2022, 2:23:38 AM
Modified	12/4/2022, 11:55:09 PM

- ## Tags:

  - #nosource
  - ⊖ No DOI found

- # Taking Down Prosecraft.io

Item Type	Web Page
Author	Benji Smith
Abstract	Today, I'm taking down the prosecraft.io website, which had previously been dedicated to the linguistic analysis of literature, including…
Date	2023-08-07T20:16:45.194Z
Language	en

URL	https://blog.shaxpir.com/taking-down-prosecraft-io-37e189797121
Accessed	8/8/2023, 2:41:55 PM
Extra	WFZ Sad.
Website Title	Medium
Date Added	8/8/2023, 2:41:55 PM
Modified	8/9/2023, 6:53:16 PM

- ## Attachments

  - ○ Snapshot

- # Team Roles & Rhetorical Intelligence in Human-Machine Writing

Item Type	Conference Paper
Author	Heidi A. McKee
Author	James E. Porter
Abstract	This paper examines AI-based writing systems and how humans might partner with these systems to produce effective professional communication. We offer a taxonomy for examining roles in human-machine teaming for writing: Resource Tool, Assistant, Writer, and Executive Decision-Maker (whether at the beginning or end of the project). In particular, we focus on humanmachine teaming in relation to what we call rhetorical intelligence, the

ability to invent and write for audience, purpose, and context. We examine human-machine writing by focusing on two cases: GameChanger and Phrazor by vPhrase. We conclude by proposing some guidelines for human-machine teaming for the production of professional communication.

Date	2022-07
Library Catalog	IEEE Xplore
Extra	Rhetorical intelligence is a useful term. Human-machine teaming less so. I love the LLMs, but you can't have a team when there is only one player.
Pages	384-391
Proceedings Title	2022 IEEE International Professional Communication Conference (ProComm)
Conference Name	2022 IEEE International Professional Communication Conference (ProComm)
DOI	10.1109/ProComm53155.2022.00078
Date Added	9/17/2022, 10:39:27 AM
Modified	9/17/2022, 10:40:38 AM

- ## Tags:

    - Uncertainty
    - Artificial intelligence
    - artificial intelligence, collaborative writing, human-computer interaction, human machine teaming, rhetoric, rhetorical intelligence

- Collaboration
- Production
- Taxonomy
- Training
- Writing

Notes:

- In particular, we focus on humanmachine teaming in relation to what we call rhetorical intelligence, t

Attachments

- IEEE Xplore Abstract Record

- # Technical Report on Neural Language Models and Few-Shot Learning for Systematic Requirements Processing in MDSE

Item Type	Preprint
Author	Vincent Bertram
Author	Miriam Boß
Author	Evgeny Kusmenko

Author	Imke Helene Nachmann
Author	Bernhard Rumpe
Author	Danilo Trotta
Author	Louis Wachtmeister

Abstract	Systems engineering, in particular in the automotive domain, needs to cope with the massively increasing numbers of requirements that arise during the development process. To guarantee a high product quality and make sure that functional safety standards such as ISO26262 are fulfilled, the exploitation of potentials of model-driven systems engineering in the form of automatic analyses, consistency checks, and tracing mechanisms is indispensable. However, the language in which requirements are written, and the tools needed to operate on them, are highly individual and require domain-specific tailoring. This hinders automated processing of requirements as well as the linking of requirements to models. Introducing formal requirement notations in existing projects leads to the challenge of translating masses of requirements and process changes on the one hand and to the necessity of the corresponding training for the requirements engineers. In this paper, based on the analysis of an open-source set of automotive requirements, we derive domain-specific language constructs helping us to avoid ambiguities in requirements and increase the level of formality. The main contribution is the adoption and evaluation of few-shot learning with large pretrained language models for the automated translation of informal requirements to structured languages such as a requirement DSL. We show that support sets of less than ten translation examples can suffice to few-shot train a language model to incorporate keywords and implement syntactic rules into informal natural language requirements.

- **Tags:**

  - Computer Science - Artificial Intelligence
  - Computer Science - Software Engineering
  - D.2.1
  - I.2.7

Attachments

  - arXiv Fulltext PDF
  - arXiv.org Snapshot

- # TEITOK: Text-faithful annotated corpora

Item Type	Conference Paper
Author	Maarten Janssen
Date	2016
Short Title	TEITOK
Library Catalog	Google Scholar
Pages	4037–4043
Proceedings Title	Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)
Date Added	12/3/2022, 11:15:41 AM
Modified	12/4/2022, 11:56:03 PM

- **Tags:**

  - #nosource
  - ⊖ No DOI found

- # TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages

Item Type	Journal Article
Author	Marti A Hearst

| Abstract | TextTiling is a technique for subdividing texts into multi-paragraph units that represent passages, or subtopics. The discourse cues for identifying major subtopic shifts are patterns of lexical co-occurrence and distribution. The algorithm is fully implemented and is shown to produce segmentation that corresponds well to human judgments of the subtopic boundaries of 12 texts. Multi-paragraph subtopic segmentation should be useful for many text analysis tasks, including information retrieval and summarization. |

Accessed	9/3/2021, 8:00:00 PM
Date Added	9/4/2021, 1:32:22 AM
Modified	12/4/2022, 11:55:07 PM

- ## Tags:

  - #nosource
  - ⊖ No DOI found

- # The 'Joan Is Awful' Episode of 'Black Mirror' Asks You to Please Click Here

Item Type	Magazine Article
Author	Amit Katwala
Abstract	Netflix's dystopian hit asks what happens when people become too content about … content.
Language	en-US

Library Catalog	www.wired.com
URL	https://www.wired.com/story/black-mirror-joan-is-awful-click-here/
Accessed	6/17/2023, 6:44:03 PM
Extra	Section: tags
Publication	Wired
ISSN	1059-1028
Date Added	6/17/2023, 6:44:03 PM
Modified	6/17/2023, 6:44:03 PM

- # Tags:

  - social media
  - streaming
  - tv

Notes:

  - viewers see in "Joan Is Awful": personalized content, generated by artificial intelligence, using deepfakes of famous actors as stand-ins for regular people. Joan—a mid-level executive at a generic startup played by Schitt's Creek's Annie Murphy—gets home after a difficult day

at the office to find the events of that very day have been turned into a slick television show starring Salma Hayek as the titular character. Things spiral from there—there are layers upon layers, content all the way down.

- # The AI feedback loop: Researchers warn of 'model collapse' as AI trains on AI-generated content

Item Type	Blog Post
Abstract	As a generative AI training model is exposed to more AI-generated data, it performs worse, producing more errors, leading to model collapse.
Date	2023-06-12T13:00:00+00:00
Language	en-US
Short Title	The AI feedback loop
URL	https://venturebeat.com/ai/the-ai-feedback-loop-researchers-warn-of-model-collapse-as-ai-trains-on-ai-generated-content/
Accessed	6/17/2023, 10:57:02 PM
Blog Title	VentureBeat
Date Added	6/17/2023, 10:57:02 PM
Modified	6/17/2023, 10:57:06 PM

- **Notes:**
    - While all this news is worrisome for current generative AI technology and the companies seeking to monetize with it, especially in the medium-to-long term, there is a silver lining for human content creators: The researchers conclude that in a future filled with gen AI tools and their content, human-created content will be even more valuable than it is today — if only as a source of pristine training data for AI.

Attachments

- Snapshot

The AI Publishing Experiment

Item Type	Web Page
Date	2021-11-08
URL	https://medium.com/the-ai-publishing-experiment
Accessed	7/19/2022, 2:46:27 PM
Extra	WFZ Author is using Jasper AI assistant to crank out 10-25,000 word books at +10/month. Uses keyword research tools to gauge market demand. While this is not a super impressive project in terms of the quality of the publications, it is an early example of truly AI-aided book creation, and shows how the tools can help author productivity.

Date Added	7/19/2022, 2:46:27 PM
Modified	7/20/2022, 8:17:22 AM

- **Attachments**

  o The AI Publishing Experiment

The Art of Dramatic Writing: Its Basis in the Creative Interpretation of Human Motives , Lajos Egri (1972)

Item Type	Journal Article
Author	Rene Rawls
Abstract	Review of: The Art of Dramatic Writing: Its Basis in the Creative Interpretation of Human Motives , Lajos Egri (1972)Originally published by Simon&Schuster in 1942 as How to Write a Play New York: Touchstone, 320 pp.,ISBN-13 978-0-67121-332-9, p/bk, $18.00
Date	2020
Volume	11
Publication	Journal of Screenwriting
DOI	10.1386/josc_00030_5
Issue	2

- **Tags:**

  - #nosource

- # The Augmented Social Scientist: Using Sequential Transfer Learning to Annotate Millions of Texts with Human-Level Accuracy

Item Type	Journal Article
Author	Salomé Do
Author	Étienne Ollion
Author	Rubing Shen
Abstract	The last decade witnessed a spectacular rise in the volume of available textual data. With this new abundance came the question of how to analyze it. In the social sciences, scholars mostly resorted to two well-established approaches, human annotation on sampled data on the one hand (either performed by the researcher, or outsourced to microworkers), and quantitative methods on the other. Each approach has its own merits - a potentially very fine-grained analysis for the former, a very scalable one for the latter - but the combination of these two properties has not yielded highly accurate results so far. Leveraging

recent advances in sequential transfer learning, we demonstrate via an experiment that an expert can train a precise, efficient automatic classifier in a very limited amount of time. We also show that, under certain conditions, expert-trained models produce better annotations than humans themselves. We demonstrate these points using a classic research question in the sociology of journalism, the rise of a ?horse race? coverage of politics. We conclude that recent advances in transfer learning help us augment ourselves when analyzing unstructured data.

Date	2022-12-04
Language	en
Short Title	The Augmented Social Scientist
Library Catalog	SAGE Journals
URL	https://doi.org/10.1177/00491241221134526
Accessed	12/8/2022, 5:27:17 AM
Extra	WFZ Good article showing how to do transfer learning at scale.
Pages	00491241221134526
Publication	Sociological Methods & Research
DOI	10.1177/00491241221134526
ISSN	0049-1241
Date Added	12/8/2022, 5:27:17 AM
Modified	12/15/2022, 8:53:39 AM

- **Tags:**
  - scalable data mining methods and algorithms

- ## The Chicago Manual of Style

Item Type	Book
Author	University of Chicago Press Editorial Staff
Abstract	Technologies may change, but the need for clear and accurate communication never goes out of style. That is why for more than one hundred years The Chicago Manual of Style has remained the definitive guide for anyone who works with words. In the seven years since the previous edition debuted, we have seen an extraordinary evolution in the way we create and share knowledge. This seventeenth edition of The Chicago Manual of Style has been prepared with an eye toward how we find, create, and cite information that readers are as likely to access from their pockets as from a bookshelf. It offers updated guidelines on electronic workflows and publication formats, tools for PDF annotation and citation management, web accessibility standards, and effective use of metadata, abstracts, and keywords. It recognizes the needs of those who are self-publishing or following open access or Creative Commons publishing models. The citation chapters reflect the ever-expanding universe of electronic sources--including social media posts and comments, private messages, and app content--and also offer updated guidelines on such issues as DOIs, time stamps, and e-book locators. Other improvements are independent of technological change. The chapter on grammar and usage includes an expanded glossary of problematic words and phrases and a new section on syntax as well as updated guidance on gender-neutral pronouns and bias-free language. Key sections on punctuation and basic citation style have been reorganized and clarified. To facilitate navigation, headings and paragraph titles have been revised and clarified throughout. And the bibliography has been updated and expanded to include the latest and best resources available. This edition continues to reflect expert insights gathered from Chicago's own staff and from an advisory board of publishing experts from across the profession. It also includes suggestions inspired by emails, calls, and even tweets from readers. No matter how much the means of communication change, The Chicago Manual of Style remains the ultimate resource for those who care about getting the details right.

Date	2017
Language	en
Library Catalog	Google Books
Extra	WFZ Essential for anyone interested in producing books.
Publisher	University of Chicago Press
ISBN	978-0-226-28705-8
# of Pages	1144
Date Added	8/3/2022, 5:17:36 PM
Modified	8/3/2022, 5:18:38 PM

- # Tags:

  - Language Arts & Disciplines / Authorship
  - LANGUAGE ARTS & DISCIPLINES / Editing & Proofreading
  - LANGUAGE ARTS & DISCIPLINES / Grammar & Punctuation
  - LANGUAGE ARTS & DISCIPLINES / Publishing
  - LANGUAGE ARTS & DISCIPLINES / Reference
  - Language Arts & Disciplines / Style Manuals
  - Reference / Writing Skills

Attachments

The Curse of Recursion: Training on Generated Data Makes Models Forget

Item Type	Preprint
Author	Ilia Shumailov
Author	Zakhar Shumaylov
Author	Yiren Zhao
Author	Yarin Gal
Author	Nicolas Papernot
Author	Ross Anderson
Abstract	Stable Diffusion revolutionised image creation from descriptive text. GPT-2, GPT-3(.5) and GPT-4 demonstrated astonishing performance across a variety of language tasks. ChatGPT introduced such language models to the general public. It is now clear that large language models (LLMs) are here to stay, and will bring about drastic change in the whole ecosystem of online text and images. In this paper we consider what the future might hold. What will happen to GPT-{n} once LLMs contribute much of the language found online? We find that use of model-generated content in training causes irreversible defects in the resulting models, where tails of the original content distribution disappear. We refer to this effect as Model Collapse and show that it can occur in Variational Autoencoders, Gaussian Mixture Models and LLMs. We build theoretical intuition behind the phenomenon and portray its ubiquity amongst all learned generative models. We demonstrate that it has to be taken seriously if we are to sustain the benefits of training from large-scale data scraped from the web.

Indeed, the value of data collected about genuine human interactions with systems will be increasingly valuable in the presence of content generated by LLMs in data crawled from the Internet.

Date	2023-05-31
Short Title	The Curse of Recursion
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2305.17493
Accessed	6/17/2023, 6:18:59 PM
Extra	arXiv:2305.17493 [cs]
DOI	10.48550/arXiv.2305.17493
Repository	arXiv
Archive ID	arXiv:2305.17493
Date Added	6/17/2023, 6:19:00 PM
Modified	6/17/2023, 6:19:00 PM

- **Tags:**

  - Computer Science - Computation and Language
  - Computer Science - Machine Learning
  - Computer Science - Artificial Intelligence

- ○ Computer Science - Computer Vision and Pattern Recognition
- ○ Computer Science - Cryptography and Security

Attachments

- ○ arXiv Fulltext PDF
- ○ arXiv.org Snapshot

• The False Promise of Imitating Proprietary LLMs

Item Type	Preprint
Author	Arnav Gudibande
Author	Eric Wallace
Author	Charlie Snell
Author	Xinyang Geng
Author	Hao Liu
Author	Pieter Abbeel
Author	Sergey Levine
Author	Dawn Song
Abstract	An emerging method to cheaply improve a weaker language model is to finetune it on outputs from a stronger model, such as a proprietary system like

ChatGPT (e.g., Alpaca, Self-Instruct, and others). This approach looks to cheaply imitate the proprietary model's capabilities using a weaker open-source model. In this work, we critically analyze this approach. We first finetune a series of LMs that imitate ChatGPT using varying base model sizes (1.5B–13B), data sources, and imitation data amounts (0.3M–150M tokens). We then evaluate the models using crowd raters and canonical NLP benchmarks. Initially, we were surprised by the output quality of our imitation models—they appear far better at following instructions, and crowd workers rate their outputs as competitive with ChatGPT. However, when conducting more targeted automatic evaluations, we find that imitation models close little to none of the gap from the base LM to ChatGPT on tasks that are not heavily supported in the imitation data. We show that these performance discrepancies may slip past human raters because imitation models are adept at mimicking ChatGPT's style but not its factuality. Overall, we conclude that model imitation is a false promise: there exists a substantial capabilities gap between open and closed LMs that, with current methods, can only be bridged using an unwieldy amount of imitation data or by using more capable base LMs. In turn, we argue that the highest leverage action for improving open-source models is to tackle the difficult challenge of developing better base LMs, rather than taking the shortcut of imitating proprietary systems.

Date	2023-05-25
Language	en
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2305.15717
Accessed	5/27/2023, 5:57:36 PM
Extra	arXiv:2305.15717 [cs]
Repository	arXiv
Archive ID	arXiv:2305.15717

Date Added	5/27/2023, 5:57:36 PM
Modified	5/27/2023, 5:57:37 PM

- # Tags:

  - Computer Science - Computation and Language

Attachments

- Gudibande et al. - 2023 - The False Promise of Imitating Proprietary LLMs.pdf

---

- # THE FUTURE IMPACT OF ARTIFICIAL INTELLIGENCE ON THE PUBLISHING INDUSTRY

Item Type	Journal Article
Author	Gould Finch
Abstract	Artificial Intelligence (AI) has always had fantastical appeal. Imaginative ideas of futuristic superhuman technology provide endless material for storytelling, inspiring humans to create worlds where the impossible is performed with ease. These ideas, however, once at home in seemingly distant dreams, are moving from fiction into reality as AI continues to take root in common business practices. While the hype surrounding advancements in AI is often confined to tech fields, many in the creative industries are taking notice as AI-specific terms like Big Data, predictive analytics, and natural language processing become household words. For those companies implementing AI in the right way at the

right time, the systems set to disrupt technology-based industries become the very tools with which they'll climb their way to the top.

Language	en
Library Catalog	Zotero
Pages	24
Date Added	7/19/2022, 7:19:47 AM
Modified	7/19/2022, 7:19:49 AM

- ## Tags:

  - ⊖ No DOI found

Attachments

  - Finch - THE FUTURE IMPACT OF ARTIFICIAL INTELLIGENCE ON TH.pdf

- # The Future of Picture Books in the Age of AI

Item Type	Web Page
Abstract	Illustrations created by artificial intelligence are a pale imitation of the real thing. They're also unethical, say illustrators.

Language	en
URL	https://www.kirkusreviews.com/news-and-features/articles/the-future-of-picture-books-in-the-age-of-ai/
Accessed	6/13/2023, 1:10:15 AM
Extra	WFZ Predictable efforts to rally opposition to exploitation, expropriation of rights. Includes a sample AI-generated illustration marked up with a devastating critique by a human artist.
Website Title	Kirkus Reviews
Date Added	6/13/2023, 1:10:14 AM
Modified	6/13/2023, 4:36:59 AM

- ## Attachments

  - ○ Snapshot

- # The history of Amazon's recommendation algorithm

Item Type	Web Page
Abstract	Collaborative filtering and beyond.
Date	2019-11-22T05:10:17.901

Language	en
URL	https://www.amazon.science/the-history-of-amazons-recommendation-algorithm
Accessed	8/27/2022, 12:03:34 AM
Extra	Section: News and features
Website Title	Amazon Science
Date Added	8/27/2022, 12:03:34 AM
Modified	8/27/2022, 12:03:43 AM

- **Attachments**
  - Snapshot

The Hollywood Actors Strike Will Revolutionize the AI Fight

Item Type	Magazine Article
Author	Angela Watercutter
Abstract	Bold-faced names like Meryl Streep and the halting of production could give artificial intelligence a whole new level of awareness.
Language	en-US

Library Catalog	www.wired.com
URL	https://www.wired.com/story/hollywood-sag-strike-artificial-intelligence/
Accessed	7/14/2023, 10:29:16 AM
Extra	WFZ Some of the world's biggest stars have generative AI in their sights.
Publication	Wired
ISSN	1059-1028
Date Added	7/14/2023, 10:29:16 AM
Modified	7/16/2023, 6:33:49 AM

- ## Tags:

  - tv
  - movies
  - film
  - hollywood
  - the monitor

Notes:

  - Will any of this stop the rise of the bots? No. It doesn't even negate that AI could be useful in a lot of fields. But what it does do is demonstrate

that people are paying attention—especially now that bold-faced names like Meryl Streep and Jennifer Lawrence are talking about artificial intelligence.

Attachments

- ○ Snapshot

- # The hundreds of changes made to Roald Dahl's books to suit a new 'sensitive' generation

Item Type	Newspaper Article
Author	Ed Cumming
Author	Genevieve Holl-Allen
Author	Benedict Smith
Abstract	Quietly and systematically, the author's entire catalogue has been reworked in a bid for 'relevancy''
Date	2023-02-17
Language	en-GB
Library Catalog	www.telegraph.co.uk

URL	https://www.telegraph.co.uk/news/2023/02/17/roald-dahl-books-rewritten-offensive-matilda-witches-twits/
Accessed	2/17/2023, 7:17:42 PM
Publication	The Telegraph
ISSN	0307-1235
Date Added	2/17/2023, 7:17:42 PM
Modified	2/17/2023, 7:17:42 PM

- **Tags:**

  - Children's books
  - Other features
  - Roald Dahl

Attachments

  - Snapshot

- # The Infinite Article

Item Type	Web Page
Author	Nathan Baschez

Abstract	How close are we to having a personal AI that can write the perfect thing just for us?
Date	2022-09-28
URL	
Accessed	9/28/2022, 8:19:24 PM
Date Added	9/28/2022, 8:19:24 PM
Modified	9/28/2022, 8:19:24 PM

- ## Attachments

  - Snapshot

The Labeled Segmentation of Printed Books

Item Type	Report
Author	Lara Mcconnaughey
Author	Jennifer Dai
Author	David Bamman
Abstract	We introduce the task of book structure labeling: segmenting and assigning a fixed category (such as TABLE OF CONTENTS, PREFACE, INDEX) to the document structure of printed books. We manually annotate the page-level structural

categories for a large dataset totaling 294,816 pages in 1,055 books evenly sampled from 1750-1922, and present empirical results comparing the performance of several classes of models. The best-performing model, a bidirectional LSTM with rich features, achieves an overall accuracy of 95.8 and a class-balanced macro F-score of 71.4.

URL	https://www.hathitrust.org/
Accessed	4/13/2021, 8:00:00 PM
Date Added	4/14/2021, 9:25:32 PM
Modified	4/14/2021, 9:28:53 PM

- **Attachments**

  - PDF

- # The Merch-ification of Book Publishing

Item Type	Web Page
Abstract	BookTok and Bookstagram have given rise to a new form of book publicity. Is it tearing the literary community apart?
Date	2023-07-11T10:00:00Z
Language	en-US
URL	https://www.esquire.com/entertainment/books/a44449495/book-publishing-influencer-merch-explained/

Accessed	7/16/2023, 5:52:11 AM
Extra	WFZ TikTok is a purely algorithmic mode forbook marketing that has grown dramatically in importance over the last few years (despite the relentless efforts of the US national security establishment to kill it and its Chinese intel sponsors dead, dead, dead). Yet as the algorithm rises in one area, the physical rises in another, as if in response. Turns out merch is the way to reach book influencers!
Website Title	Esquire
Date Added	7/16/2023, 5:52:11 AM
Modified	7/16/2023, 6:29:48 AM

- ## **Attachments**
  - ○ Snapshot

- # The Mueller Report's Free, but Publishers Are Selling It as a Book

Item Type	Newspaper Article
Author	Laura M. Holson
Abstract	A number of book publishers are selling the report, released on Thursday, even though it can be read for free online.
Date	2019-04-18

Language	en-US
Library Catalog	NYTimes.com
URL	https://www.nytimes.com/2019/04/18/business/download-mueller-report-buy.html
Accessed	6/19/2022, 2:10:26 AM
Extra	WFZ Prominent albeit largely routine review of "instant publishing" of government public documents, but was especially notable to me because Nimble Books is mentioned for its reprint of the Pentagon Papers. I'm experimenting with various methods of annotating such documents.
Section	Business
Publication	The New York Times
ISSN	0362-4331
Date Added	6/19/2022, 2:10:26 AM
Modified	8/3/2022, 8:02:14 AM

- ## Tags:

  - Books and Literature
  - Mueller, Robert S III
  - Russian Interference in 2016 US Elections and Ties to Trump Associates
  - Trump, Donald J

Notes:

○ In 1998, for example, The Post and the publisher PublicAffairs released a paperback version of "The Starr Report: The Findings of Independent Counsel Kenneth W. Starr on President Clinton and the Lewinsky Affair." And in 2011, the National Archives and Nimble Books published "The Pentagon Papers," the report commissioned by Secretary of Defense Robert S. McNamara in 1967 that led to the end of the Vietnam War.

Attachments

○ Snapshot

The Old Guard Is Out at Penguin Random House

Item Type	Web Page
Author	Shawn McCreesh
Abstract	A generational shift means a prestige purge.
Date	2023-07-18T10:00:39.319-04:00
Language	en-us

URL	https://nymag.com/intelligencer/2023/07/the-old-guard-is-out-at-penguin-random-house.html
Accessed	7/20/2023, 2:55:04 AM
Extra	WFZ Paywall ($). Money quote: "A whole generation of editors has been tapping its foot, waiting for these boomers to clear out. "Knopf in particular has stables of people, almost like The New Yorker," says the youngster. "It's like, How do you even work here still?" Well, they won't for long." The irony is that all these youngsters will be replaced by AIs before they know what hit them.
Website Title	Intelligencer
Date Added	7/20/2023, 2:55:04 AM
Modified	7/20/2023, 2:57:44 AM

- ## Attachments

  - ○ Snapshot

- # The Perils of Using Mechanical Turk to Evaluate Open-Ended Text Generation

Item Type	Preprint
Author	Marzena Karpinska
Author	Nader Akoury

Author	Mohit Iyyer

Abstract	Recent text generation research has increasingly focused on open-ended domains such as story and poetry generation. Because models built for such tasks are difficult to evaluate automatically, most researchers in the space justify their modeling choices by collecting crowdsourced human judgments of text quality (e.g., Likert scores of coherence or grammaticality) from Amazon Mechanical Turk (AMT). In this paper, we first conduct a survey of 45 open-ended text generation papers and find that the vast majority of them fail to report crucial details about their AMT tasks, hindering reproducibility. We then run a series of story evaluation experiments with both AMT workers and English teachers and discover that even with strict qualification filters, AMT workers (unlike teachers) fail to distinguish between model-generated text and human-generated references. We show that AMT worker judgments improve when they are shown model-generated output alongside human-generated references, which enables the workers to better calibrate their ratings. Finally, interviews with the English teachers provide deeper insights into the challenges of the evaluation process, particularly when rating model-generated text.

Date	2021-09-14
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2109.06835
Accessed	8/10/2022, 12:06:31 AM
Extra	WFZ Testing & evaluation is a major problem in longform generation. The paper finds that most researchers rely on Amazon Mechanical Turk or similar out-sourcing methods and that unsurprisingly having gig workers evaluate writing quality can be problematic. This is an area where I think I bring a comparative advantage to this work: AI cannot yet match a highly experienced and discerning editorial taste. But the day is coming when an AI can match the power of a lifetime of reading, thinking, writing, and lived experience.
DOI	10.48550/arXiv.2109.06835
Repository	arXiv

Archive ID	arXiv:2109.06835
Date Added	8/10/2022, 12:06:31 AM
Modified	8/10/2022, 12:11:05 AM

- ## Tags:

  - Computer Science - Computation and Language

Notes:

  - Comment: EMNLP 2021 (20 pages)

Attachments

  - arXiv Fulltext PDF
  - arXiv.org Snapshot

- # The Practical Guides for Large Language Models

Item Type	Software
Programmer	Mooler0410
Date	2023-04-28T02:07:43Z

Library Catalog	GitHub
URL	https://github.com/Mooler0410/LLMsPracticalGuide
Accessed	4/27/2023, 10:09:44 PM
Extra	WFZ includes excellent genealogical tree of LLMs. <img src="https://github.com/Mooler0410/LLMsPracticalGuide/raw/main/imgs/survey-gif-test.gif">
Date Added	4/27/2023, 10:09:44 PM
Modified	4/27/2023, 10:21:03 PM

- # The Royal Reading #midjourney #aiartwork https://t.co/8ShFIULNLr

Item Type	Artwork
Artist	Destiny K ⬚⬚⬚ [@destinykrainbow]
Abstract	A painting of two royal women reading atop a pile of leather bound books. https://pbs.twimg.com/media/FkPNrSNXgAAxV5v?format=jpg&name=medium
Date	2022-12-18T05:45Z
Language	en
URL	https://twitter.com/destinykrainbow/status/1604352040494301185
Accessed	12/19/2022, 2:59:32 AM

Extra	WFZ Generative image AI being used to make book-reading look very prestigious and very cool. I approve.
Date Added	12/19/2022, 2:59:32 AM
Modified	12/19/2022, 2:41:23 PM

- # Notes:

  - o See new Tweets Conversation Destiny K @destinykrainbow The Royal Reading #midjourney #aiartwork

Attachments

  - o Link

- # The State Internet Information Office and other three departments issued the "Regulations on the Administration of Deep Synthesis of Internet Information Services"-Office of the Central Committee of the Communist Party of China

Item Type	Web Page

Author	Office of the Cyber Security and Informatization Commission of the Central Committee of the Communist Party of China
Date	2022-12-11
URL	http://www.cac.gov.cn/2022-12/11/c_1672221949318230.htm
Accessed	12/13/2022, 3:19:38 AM
Extra	WFZ Press release from PRC govt.
Date Added	12/13/2022, 3:19:38 AM
Modified	12/15/2022, 8:46:24 AM

- ## **Tags:**

  - law \& legislation

- # The state of the art in map-like visualization

Item Type	Conference Paper
Author	Marius Hogräfer
Author	Magnus Heitzler
Author	Hans-Jörg Schulz

Date	2020
Extra	WFZ I have a long-standing affection for fantasy maps of literature. This is adjacent.
Volume	39
Pages	647–674
Proceedings Title	Computer graphics forum
Date Added	12/26/2022, 10:28:55 AM
Modified	12/30/2022, 8:30:54 AM

- ## Tags:

  - ⊖ No DOI found

- # The Story Behind all the Animals on O'Reilly's Book Covers

Item Type	Blog Post
Author	Cassell, David
Abstract	For nearly three decades the animals on the O'Reilly boo covers have become almost a part of geek culture itself. The woman behind this initiative? Edie Freedman, who worked as a creative director for the publishing company.
Date	2019-07-07T13:00:58+00:00

Language	en-US
URL	https://thenewstack.io/the-story-behind-all-the-animals-on-oreillys-book-covers/
Accessed	8/6/2022, 3:39:55 AM
Extra	WFZ The sketches were initially drawn from the Dover Pictorial Archives.
Blog Title	The New Stack
Date Added	8/6/2022, 3:39:55 AM
Modified	8/6/2022, 5:00:44 AM

- ## **Attachments**
  - ○ Snapshot

- # The Vanity Fair diaries: 1983-1992

Item Type	Book
Author	Tina Brown
Abstract	The irreverent diaries of the author's celebrated years as editor-in-chief of Vanity Fair Condé Nast's troubled periodical and her experiences within the cutthroat world of gla
Date	2017
Language	eng

Short Title	The Vanity Fair diaries
Library Catalog	Open WorldCat
URL	http://firstsearch.oclc.org/WebZ/DECRead?standardNoType=1&standardNo=162779
Accessed	9/17/2022, 12:31:38 AM
Extra	WFZ Tina Brown's diaries are one of the better windows I have seen into the mind of least as good as Tina Brown was, and at least as transparent about its thinking as she i
Place	New York
Publisher	Henry Holt and Company
ISBN	978-1-62779-136-6
Edition	First edition
# of Pages	436
Date Added	9/17/2022, 12:31:39 AM
Modified	12/4/2022, 11:55:45 PM

- # Tags:

  - United States
  - #nosource
  - Autobiographies
  - Biographies

- # The Web Is Your Oyster - Knowledge-Intensive NLP against a Very Large Web Corpus

Item Type	Journal Article
Author	Aleksandra Piktus
Author	Fabio Petroni
Author	Vladimir Karpukhin
Author	Dmytro Okhonko

Author   Samuel Broscheit

Author   Gautier Izacard

Author   Patrick Lewis

Author   Barlas Oğuz

Author   Edouard Grave

Author   Wen-tau Yih

Author   Sebastian Riedel

Abstract   In order to address increasing demands of real-world applications, the research for knowledge-intensive NLP (KI-NLP) should advance by capturing the challenges of a truly open-domain environment: web-scale knowledge, lack of structure, inconsistent quality and noise. To this end, we propose a new setup for evaluating existing knowledge intensive tasks in which we generalize the background corpus to a universal web snapshot. We investigate a slate of NLP tasks which rely on knowledge - either factual or common sense, and ask systems to use a subset of CCNet - the Sphere corpus - as a knowledge source. In contrast to Wikipedia, otherwise a common background corpus in KI-NLP, Sphere is orders of magnitude larger and better reflects the full diversity of knowledge on the web. Despite potential gaps in coverage, challenges of scale, lack of structure and lower quality, we find that retrieval from Sphere enables a state of the art system to match and even outperform Wikipedia-based models on several tasks. We also observe that while a dense index can outperform a sparse BM25 baseline on Wikipedia, on Sphere this is not yet possible. To facilitate further research and minimise the community's reliance on proprietary, black-box search engines, we share our indices, evaluation metrics and infrastructure.

Date   2021

- ## Tags:

  - Artificial Intelligence (cs.AI)
  - Computation and Language (cs.CL)
  - FOS: Computer and information sciences
  - Information Retrieval (cs.IR)
  - Machine Learning (cs.LG)

Attachments

  - [2112.09924] The Web Is Your Oyster - Knowledge-In.pdf

- # The WEIRDest People in the World: How the West Became Psychologically Peculiar and Particularly Prosperous

Item Type	Book
Author	Joseph Henrich
Abstract	The book begins with a detailed examination of the introduction of Gutenberg's printed book and its profound impact on Western psychology.
Date	2020-09-08
Language	English
Short Title	The WEIRDest People in the World
Library Catalog	Amazon
Publisher	Farrar, Straus and Giroux
# of Pages	706
Date Added	6/15/2022, 6:48:36 PM
Modified	12/4/2022, 11:55:11 PM

- # Tags:

  - #nosource

- # The Why, What and How of Artificial General Intelligence Chip Development

Item Type	Report
Author	Alex James
Abstract	The AI chips increasingly focus on implementing neural computing at low power and cost. The intelligent sensing, automation, and edge computing applications have been the market drivers for AI chips. Increasingly, the generalisation, performance, robustness, and scalability of the AI chip solutions are compared with human-like intelligence abilities. Such a requirement to transit from application-specific to general intelligence AI chip must consider several factors. This paper provides an overview of this cross-disciplinary field of study, elaborating on the generalisation of intelligence as understood in building artificial general intelligence (AGI) systems. This work presents a listing of emerging AI chip technologies, classification of edge AI implementations, and the funnel design flow for AGI chip development. Finally, the design consideration required for building an AGI chip is listed along with the methods for testing and validating it.
URL	https://www.iiitmk.ac.in/agins/apj/
Accessed	3/28/2021, 8:00:00 PM
Extra	arXiv: 2012.06338v1 ISBN: 2012.06338v1
Date Added	3/29/2021, 11:01:26 PM

- **Tags:**
  - AI chips
  - AI hardware
  - Edge AI
  - Index Terms-Artificial General Intelligence
  - #nosource

- # The World's Technological Capacity to Store, Communicate, and Compute Information

Item Type	Journal Article
Author	Martin Hilbert
Author	Priscila López
Abstract	An inventory of the world's technological capacity from 1986 to 2007 reveals the evolution from analog to digital technologies. , We estimated the world's technological capacity to store, communicate, and compute information, tracking 60 analog and digital technologies during the period from 1986 to 2007. In 2007, humankind was able to store 2.9×10^{20} optimally compressed bytes, communicate almost 2×10^{21} bytes, and carry out 6.4×10^{18} instructions per second on general-purpose computers. General-purpose computing capacity grew at an annual rate of 58%. The world's capacity for bidirectional telecommunication grew at 28% per year, closely followed by the increase in globally stored information (23%). Humankind's capacity for unidirectional information diffusion through broadcasting channels has experienced comparatively modest annual growth (6%). Telecommunication has been dominated by digital technologies since 1990 (99.9% in digital format

in 2007), and the majority of our technological memory has been in digital format since the early 2000s (94% digital in 2007).

Date	04/2011
Language	en
Library Catalog	DOI.org (Crossref)
URL	https://www.science.org/doi/10.1126/science.1200970
Accessed	8/3/2022, 12:47:44 AM
Extra	WFZ Useful reference points for comparison to the global book corpus.
Volume	332
Pages	60-65
Publication	Science
DOI	10.1126/science.1200970
Issue	6025
Journal Abbr	Science
ISSN	0036-8075, 1095-9203
Date Added	8/3/2022, 12:47:44 AM
Modified	12/4/2022, 11:55:38 PM

- **Tags:**

  - #nosource

- # TinyStories: How Small Can Language Models Be and Still Speak Coherent English?

Item Type	Preprint
Author	Ronen Eldan
Author	Yuanzhi Li
Abstract	Language models (LMs) are powerful tools for natural language processing, but they often struggle to produce coherent and fluent text when they are small. Models with around 125M parameters such as GPT-Neo (small) or GPT-2 (small) can rarely generate coherent and consistent English text beyond a few words even after extensive training. This raises the question of whether the emergence of the ability to produce coherent English text only occurs at larger scales (with hundreds of millions of parameters or more) and complex architectures (with many layers of global attention). In this work, we introduce TinyStories, a synthetic dataset of short stories that only contain words that a typical 3 to 4-year-olds usually understand, generated by GPT-3.5 and GPT-4. We show that TinyStories can be used to train and evaluate LMs that are much smaller than the state-of-the-art models (below 10 million total parameters), or have much simpler architectures (with only one transformer block), yet still produce fluent and consistent stories with several paragraphs that are diverse and have almost perfect grammar, and demonstrate reasoning capabilities. We also introduce a new paradigm for the evaluation of language models: We suggest a framework which uses GPT-4 to grade the content generated by these models as if those were stories written by students and graded by a (human) teacher. This new paradigm overcomes the flaws of standard benchmarks which often requires the model's output to be very structures, and moreover provides a multidimensional score for the model, providing scores for different capabilities such as grammar, creativity and consistency. We hope that TinyStories can facilitate the development, analysis and research of LMs, especially for low-

resource or specialized domains, and shed light on the emergence of language capabilities in LMs.

Date	2023-05-24
Short Title	TinyStories
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2305.07759
Accessed	6/25/2023, 11:18:30 PM
Extra	arXiv:2305.07759 [cs]
DOI	10.48550/arXiv.2305.07759
Repository	arXiv
Archive ID	arXiv:2305.07759
Date Added	6/25/2023, 11:18:30 PM
Modified	6/25/2023, 11:18:30 PM

- # Tags:
  - Computer Science - Computation and Language
  - Computer Science - Machine Learning
  - Computer Science - Artificial Intelligence

Attachments

- o **arXiv Fulltext PDF**
- o **arXiv.org Snapshot**

- # Towards Individuated Reading Experiences: Different Fonts Increase Reading Speed for Different Individuals

Item Type	Journal Article
Author	Shaun Wallace
Author	Zoya Bylinskii
Author	Jonathan Dobres
Author	Bernard Kerr
Author	Sam Berlow
Author	Rick Treitman
Author	Nirmal Kumawat
Author	Kathleen Arpin
Author	Dave B. Miller

Author	Jeff Huang
Author	Ben D. Sawyer
Abstract	In our age of ubiquitous digital displays, adults often read in short, opportunistic interludes. In this context of Interlude Reading, we consider if manipulating font choice can improve adult readers' reading outcomes. Our studies normalize font size by human perception and use hundreds of crowdsourced participants to provide a foundation for understanding, which fonts people prefer and which fonts make them more effective readers. Participants' reading speeds (measured in words-per-minute (WPM)) increased by 35% when comparing fastest and slowest fonts without affecting reading comprehension. High WPM variability across fonts suggests that one font does not fit all. We provide font recommendations related to higher reading speed and discuss the need for individuation, allowing digital devices to match their readers' needs in the moment. We provide recommendations from one of the most significant online reading efforts to date. To complement this, we release our materials and tools with this article.
Date	March 31, 2022
Short Title	Towards Individuated Reading Experiences
Library Catalog	ACM Digital Library
URL	https://dl.acm.org/doi/10.1145/3502222
Accessed	6/27/2023, 6:56:01 PM
Volume	29
Pages	38:1–38:56
Publication	ACM Transactions on Computer-Human Interaction
DOI	10.1145/3502222

Issue	4
Journal Abbr	ACM Trans. Comput.-Hum. Interact.
ISSN	1073-0516
Date Added	6/27/2023, 6:56:01 PM
Modified	6/27/2023, 6:56:01 PM

- **Tags:**

  - text
  - crowdsourcing
  - individuation
  - Reading
  - typography

Attachments

  - Full Text PDF

- # Towards More Effective Text Summarization Based on Textual Association Networks

Item Type	Journal Article

Author	Yuhui Tao
Author	Shuigeng Zhou
Author	Wai Lam
Author	Jihong Guan
Extra	Citation Key: Tao ISBN: 9780769534015
Pages	235-240
DOI	10/cdpqvw
Date Added	3/29/2021, 6:51:42 PM
Modified	12/4/2022, 11:52:26 PM

- **Tags:**

  - #nosource

- # Tragic and Comical Networks. Clustering Dramatic Genres According to Structural Properties

Item Type	Preprint
Author	Szemes Botond

Author	Vida Bence
Abstract	There is a growing tradition in the joint field of network studies and drama history that produces interpretations from the character networks of the plays.The potential of such an interpretation is that the diagrams provide a different representation of the relationships between characters as compared to reading the text or watching the performance. Our aim is to create a method that is able to cluster texts with similar structures on the basis of the play's well-interpretable and simple properties, independent from the number of characters in the drama, or in other words, the size of the network. Finding these features is the most important part of our research, as well as establishing the appropriate statistical procedure to calculate the similarities between the texts. Our data was downloaded from the DraCor database and analyzed in R (we use the GerDracor and the ShakeDraCor sub-collection). We want to propose a robust method based on the distribution of words among characters; distribution of characters in scenes, average length of speech acts, or character-specific and macro-level network properties such as clusterization coefficient and network density. Based on these metrics a supervised classification procedure is applied to the sub-collections to classify comedies and tragedies using the Support Vector Machine (SVM) method. Our research shows that this approach can also produce reliable results on a small sample size.
Date	2023-02-16
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2302.08258
Accessed	2/20/2023, 3:15:52 PM
Extra	arXiv:2302.08258 [cs]
Repository	arXiv
Archive ID	arXiv:2302.08258
Date Added	2/20/2023, 3:15:52 PM

Modified	2/20/2023, 3:15:52 PM

- **Tags:**

  - Computer Science - Computation and Language

Attachments

  - **arXiv Fulltext PDF**
  - **arXiv.org Snapshot**

- # Training language models to follow instructions with human feedback

Item Type	Journal Article
Author	Long Ouyang
Author	Jeff Wu
Author	Xu Jiang
Author	Diogo Almeida
Author	Carroll L. Wainwright
Author	Pamela Mishkin
Author	Chong Zhang

Author   Sandhini Agarwal

Author   Katarina Slama

Author   Alex Ray

Author   John Schulman

Author   Jacob Hilton

Author   Fraser Kelton

Author   Luke Miller

Author   Maddie Simens

Author   Amanda Askell

Author   Peter Welinder

Author   Paul Christiano

Author   Jan Leike

Author   Ryan Lowe

Abstract   Making language models bigger does not inherently make them better at following a user's intent. For example, large language models can generate outputs that are untruthful, toxic, or simply not helpful to the user. In other words, these models are not aligned with their users. In this paper, we show an avenue for aligning language models with user intent on a wide range of tasks by fine-tuning with human feedback. Starting with a set of labeler-written prompts and prompts submitted through the OpenAI API, we collect a dataset of labeler demonstrations of the desired model behavior, which we use to fine-tune GPT-3 using supervised learning. We then collect a dataset of rankings of model outputs, which we use to further fine-tune this supervised model using

reinforcement learning from human feedback. We call the resulting models InstructGPT. In human evaluations on our prompt distribution, outputs from the 1.3B parameter InstructGPT model are preferred to outputs from the 175B GPT-3, despite having 100x fewer parameters. Moreover, InstructGPT models show improvements in truthfulness and reductions in toxic output generation while having minimal performance regressions on public NLP datasets. Even though InstructGPT still makes simple mistakes, our results show that fine-tuning with human feedback is a promising direction for aligning language models with human intent.

Date	2022-03-04
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2203.02155
Accessed	5/8/2022, 10:59:22 PM
Extra	WFZ This is the definitive article describing OpenAI's Instruct model, which is a major advance that changes interaction with the AI from "do as I do" to "do this like that". This ability to interpret explicit instructions makes the system more responsive to user needs.
Publication	arXiv:2203.02155 [cs]
Date Added	5/8/2022, 10:59:22 PM
Modified	12/4/2022, 11:55:11 PM

- # Tags:

  - #nosource
  - ⊖ No DOI found
  - Computer Science - Computation and Language

- # Transcending Scaling Laws with 0.1% Extra Compute

Item Type	Preprint
Author	Yi Tay
Author	Jason Wei
Author	Hyung Won Chung
Author	Vinh Q. Tran
Author	David R. So
Author	Siamak Shakeri
Author	Xavier Garcia
Author	Huaixiu Steven Zheng
Author	Jinfeng Rao
Author	Aakanksha Chowdhery
Author	Denny Zhou
Author	Donald Metzler

Author	Slav Petrov
Author	Neil Houlsby
Author	Quoc V. Le
Author	Mostafa Dehghani

Abstract Scaling language models improves performance but comes with significant computational costs. This paper proposes UL2R, a method that substantially improves existing language models and their scaling curves with a relatively tiny amount of extra compute. The key idea is to continue training a state-of-the-art large language model (e.g., PaLM) on a few more steps with UL2's mixture-of-denoiser objective. We show that, with almost negligible extra computational costs and no new sources of data, we are able to substantially improve the scaling properties of large language models on downstream metrics. In this paper, we continue training PaLM with UL2R, introducing a new set of models at 8B, 62B, and 540B scale which we call U-PaLM. Impressively, at 540B scale, we show an approximately 2x computational savings rate where U-PaLM achieves the same performance as the final PaLM 540B model at around half its computational budget (i.e., saving $\sim$4.4 million TPUv4 hours). We further show that this improved scaling curve leads to 'emergent abilities' on challenging BIG-Bench tasks -- for instance, U-PaLM does much better than PaLM on some tasks or demonstrates better quality at much smaller scale (62B as opposed to 540B). Overall, we show that U-PaLM outperforms PaLM on many few-shot setups, i.e., English NLP tasks (e.g., commonsense reasoning, question answering), reasoning tasks with chain-of-thought (e.g., GSM8K), multilingual tasks (MGSM, TydiQA), MMLU and challenging BIG-Bench tasks. Finally, we provide qualitative examples showing the new capabilities of U-PaLM for single and multi-span infilling.

Date	2022-10-20
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2210.11399
Accessed	10/25/2022, 5:13:16 AM

Extra	WFZ Scaling is a critical issue for long-form generation. Do the techniques described here make it easier to scale the size of generations?
DOI	10.48550/arXiv.2210.11399
Repository	arXiv
Archive ID	arXiv:2210.11399
Date Added	10/25/2022, 5:13:17 AM
Modified	10/26/2022, 2:56:21 PM

- ## Tags:

  - Computer Science - Computation and Language
  - Computer Science - Machine Learning
  - Computer Science - Artificial Intelligence

Attachments

  - arXiv Fulltext PDF
  - arXiv.org Snapshot

- # Travis Fischer on Twitter

Item Type	Web Page
Abstract	"(GPT-3.5 Turbo appears to be referring to the ChatGPT Turbo model) Expanded product brief and full source: https://t.co/GhBlSOR0ZA"

Date	2023-02-21
Language	en
URL	https://twitter.com/transitive_bs/status/1628118176524533760
Accessed	2/22/2023, 4:58:36 PM
Extra	WFZ Has link to image of OpenAI's Foundry product offering which gives dedicated access to their servers.$264K/year is price of entry.
Website Title	Twitter
Date Added	2/22/2023, 4:58:36 PM
Modified	2/26/2023, 1:01:30 AM

- **Attachments**
  - ○ Snapshot

- # tstramer/midjourney-diffusion – Run with an API on Replicate

Item Type	Web Page
Author	Tal Stramer
URL	https://replicate.com/tstramer/midjourney-diffusion

Accessed	1/24/2023, 11:11:40 PM
Extra	WFZ Clean Python access to Midjourney-like capabiltiies.
Website Title	tstramer/midjourney-diffusion
Date Added	1/24/2023, 11:11:40 PM
Modified	1/26/2023, 2:12:06 PM

- ## **Attachments**

  - tstramer/midjourney-diffusion – Run with an API on Replicate

- # Ubisoft Proudly Announces 'AI' Is Helping Write Dialogue

Item Type	Web Page
Abstract	Ubisoft Ghostwriter is described by the company as 'an AI tool'
Date	2023-03-21T22:20:00.316Z
Language	en
URL	https://kotaku.com/ubisoft-ai-writing-scriptwriting-ghostwriter-machine-1850250316

Accessed	3/22/2023, 3:45:08 PM
Website Title	Kotaku
Date Added	3/22/2023, 3:45:08 PM
Modified	3/22/2023, 3:45:11 PM

- **Attachments**
  - Snapshot

- # Uncovering the Narrative Structure of "Breaking Bad" Through a Multi-Dimensional Quantitative Analysis

Item Type	Journal Article
Author	Maurizio Naldi
Author	Paola Dalla Torre
Date	2022
Library Catalog	Google Scholar
Volume	8

Pages	17–32
Publication	Series-International Journal of TV Serial Narratives
Issue	1
Date Added	8/5/2022, 8:53:34 PM
Modified	10/19/2022, 11:08:05 PM

- ## Tags:

  - ⊖ No DOI found

Attachments

  - Full Text
  - Snapshot

- # Universal and Transferable Attacks on Aligned Language Models

Item Type	Web Page
URL	https://llm-attacks.org/?utm_source=newsletter&utm_medium=email&utm_campaign=newsletter_ax
Accessed	7/30/2023, 4:14:47 PM

| Date Added | 7/30/2023, 4:14:47 PM |
| Modified | 7/30/2023, 4:14:50 PM |

- **Attachments**

  - Universal and Transferable Attacks on Aligned Language Models

---

- # Unsupervised statistical text simplification using pre-trained language modeling for initialization

Item Type	Journal Article
Author	Jipeng Qiang
Author	Feng Zhang
Author	Yun Li
Author	Yunhao Yuan
Author	Yi Zhu
Author	Xin Wu
Date	2022

Volume	17
Pages	1-10
Publication	Frontiers of Computer Science
Date Added	9/19/2022, 3:31:24 PM
Modified	9/23/2022, 3:24:51 PM

- **Tags:**

  - ⊖ No DOI found

Attachments

  - Unsupervised statistical text simplification using pre-trained language modeling for initialization | Semantic Scholar

- # Use of ChatGPT generated text for content on Stack Overflow is temporarily banned.

Item Type	Journal Article
Author	Stack Overflow

Abstract	The volume of these answers (thousands) and the fact that the answers often require a detailed read by someone with at least some subject matter expertise in order to determine that the answer is actually bad has effectively swamped our volunteer-based quality curation infrastructure. As such, we need the volume of these posts to reduce and we need to be able to deal with the ones which are posted quickly, which means dealing with users, rather than individual posts. So, for now, the use of ChatGPT to create posts here on Stack Overflow is not permitted.
Date	2022-12-05
Language	English
URL	https://meta.stackoverflow.com/questions/421831/temporary-policy-chatgpt-is-banned
Extra	WFZ StackOverflow's community moderators were quickly overwhelmed by ChatGPT posts and SO had to implement a flat ban. When ppl figure out how to use ChatGPT to create book-length manuscripts, literary agents, editorial assistants and acquisition editors will take their turn in the barrel.
Publication	Stack Overflow (Meta)
Date Added	12/5/2022, 11:15:26 PM
Modified	12/5/2022, 11:21:04 PM

- # Using Large Language Models to Simulate Multiple Humans

Item Type	Preprint
Author	Gati Aher

Author	Rosa I. Arriaga
Author	Adam Tauman Kalai

| Abstract | We propose a method for using a large language model, such as GPT-3, to simulate responses of different humans in a given context. We test our method by attempting to reproduce well-established economic, psycholinguistic, and social experiments. The method requires prompt templates for each experiment. Simulations are run by varying the (hypothetical) subject details, such as name, and analyzing the text generated by the language model. To validate our methodology, we use GPT-3 to simulate the Ultimatum Game, garden path sentences, risk aversion, and the Milgram Shock experiments. In order to address concerns of exposure to these studies in training data, we also evaluate simulations on novel variants of these studies. We show that it is possible to simulate responses of different people and that their responses are largely consistent with prior human studies from the literature. Using large language models as simulators offers advantages but also poses risks. Our use of a language model for simulation is contrasted with anthropomorphic views of a language model as having its own behavior. |

Date	2022-09-02
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2208.10264
Accessed	9/13/2022, 3:44:00 AM
Extra	arXiv:2208.10264 [cs]
DOI	10.48550/arXiv.2208.10264
Repository	arXiv
Archive ID	arXiv:2208.10264

Date Added	9/13/2022, 3:44:00 AM
Modified	9/13/2022, 3:44:03 AM

- # Tags:
  - Computer Science - Computation and Language
  - Computer Science - Machine Learning
  - Computer Science - Artificial Intelligence

Notes:

  - Comment: Added additional risks and acknowledgments

Attachments

  - arXiv Fulltext PDF
  - arXiv.org Snapshot

viewcontent.pdf

Item Type	Attachment
URL	https://bulletinofcas.researchcommons.org/cgi/viewcontent.cgi?article=2047&context=j...
Accessed	8/22/2022, 4:17:28 PM

Date Added	8/22/2022, 4:17:28 PM
Modified	12/4/2022, 11:55:42 PM

- **Tags:**

  - #nosource

- # Visual Evaluation of Text Features for Document Summarization and Analysis

Item Type	Journal Article
Author	Daniela Oelke
Author	Peter Bak
Author	Daniel A Keim
Author	Mark Last
Extra	Citation Key: Oelke ISBN: 9781424429356
Pages	75-82
Date Added	3/29/2021, 6:50:32 PM

Modified	12/4/2022, 11:51:01 PM

- ## Tags:

  - #nosource

- # Visualizing and Interacting with Geospatial Networks: A Survey and Design Space

Item Type	Journal Article
Author	Sarah Schöttler
Author	Yalong Yang
Author	Hanspeter Pfister
Author	Benjamin Bach
Abstract	This paper surveys visualization and interaction techniques for geospatial networks from a total of 95 papers. Geospatial networks are graphs where nodes and links can be associated with geographic locations. Examples can include social networks, trade and migration, as well as traffic and transport networks. Visualizing geospatial networks poses numerous challenges around the integration of both network and geographical information as well as additional information such as node and link attributes, time and uncertainty. Our overview analyses existing techniques along four dimensions: (i) the representation of geographical information, (ii) the representation of network information, (iii) the visual integration of both and (iv) the use of interaction. These four dimensions allow us to discuss techniques with respect to the trade-offs they make between showing information across all these dimensions and

how they solve the problem of showing as much information as necessary while maintaining readability of the visualization. https://geonetworks.github.io.

Date	2021
Language	en
Short Title	Visualizing and Interacting with Geospatial Networks
Library Catalog	Wiley Online Library
URL	https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14198
Accessed	12/26/2022, 12:27:42 AM
Extra	_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14198
Volume	40
Pages	5-33
Publication	Computer Graphics Forum
DOI	10.1111/cgf.14198
Issue	6
ISSN	1467-8659
Date Added	12/26/2022, 12:27:42 AM
Modified	12/26/2022, 12:27:42 AM

- **Attachments**

  - Full Text PDF
  - Snapshot
  - Visualizing and Interacting with Geospatial Networks: A Survey and Design Space

Watch Fans Have Eyes on Geneva

Item Type	Newspaper Article
Author	Robin Swithinbank
Abstract	TAG Heuer, Zenith and IWC are just a few of the brands introducing new models at the annual Watches and Wonders show.
Date	2023-03-27
Language	en-US
Library Catalog	NYTimes.com
URL	https://www.nytimes.com/2023/03/27/fashion/watch-tag-heuer-zenith-iwc-geneva.html
Accessed	3/27/2023, 4:01:29 AM
Section	Fashion
Publication	The New York Times
ISSN	0362-4331

Date Added	3/27/2023, 4:01:29 AM
Modified	3/27/2023, 4:01:29 AM

- # Tags:

  - Audemars Piguet Group
  - Breitling SA
  - Cartier SA
  - Conventions, Fairs and Trade Shows
  - Geneva (Switzerland)
  - Hermes International SA
  - Luxury Goods and Services
  - Omega SA
  - Patek Philippe SA
  - Swatch Group
  - Watches and Clocks
  - Watches and Wonders
  - Watches and Wonders Geneva Foundation

Attachments

  - Snapshot

- # Wayback Machine

Item Type	Web Page

- **Attachments**

  - PDF Snapshot

- # We Asked GPT-3 to Write an Academic Paper about Itself—Then We Tried to Get It Published

Item Type	Web Page
Author	Almira Osmanovic Thunström
Abstract	An artificially intelligent first author presents many ethical questions—and could upend the publishing process
Language	en

URL	https://www.scientificamerican.com/article/we-asked-gpt-3-to-write-an-academic-paper-about-itself-then-we-tried-to-get-it-published/
Accessed	6/30/2022, 6:38:22 PM
Website Title	Scientific American
Date Added	6/30/2022, 6:38:22 PM
Modified	6/30/2022, 6:38:28 PM

- ## Attachments

  - Snapshot

- # We Need to Talk About the Backlist

Item Type	Web Page
Author	Thad McIlroy
Abstract	Backlist matters for two simple and connected reasons: it's two thirds of what people buy, and it's markedly more profitable for publishers, with better margins and fewer returns from bookstores. Period.
Date	2021-05-24T06:42:17+00:00
Language	en
URL	https://thefutureofpublishing.com/2021/05/we-need-to-talk-about-the-backlist/

Accessed	11/11/2022, 12:02:35 AM
Extra	WFZ Well researched survey of history and financial fundamentals of backlist, arriving at the unsurprising conclusion that it pays to refresh your metadata.
Website Title	Thad McIlroy - Future Of Publishing
Date Added	11/11/2022, 12:02:34 AM
Modified	11/11/2022, 7:57:52 PM

- ## **Attachments**

  - ○ Snapshot

- # Welcome to the Age of Citizen Data Scientists

Item Type	Web Page
Author	Thuwarakesh Murallie
Abstract	Data science has been democratized for the most part. AI is now mainstream!
Date	2021-11-10T10:50:32.089Z
Language	en
URL	https://towardsdatascience.com/how-to-become-a-citizen-data-scientist-294660da0494

Accessed	8/24/2022, 9:34:35 PM
Website Title	Medium
Date Added	8/24/2022, 9:34:35 PM
Modified	8/24/2022, 9:38:36 PM

- ## Tags:

  - ETL

Attachments

  - Snapshot

- # WeLM: A Well-Read Pre-trained Language Model for Chinese

Item Type	Preprint
Author	Hui Su
Author	Xiao Zhou
Author	Houjing Yu
Author	Yuwen Chen

Author	Zilin Zhu
Author	Yang Yu
Author	Jie Zhou
Abstract	Large Language Models pre-trained with self-supervised learning have demonstrated impressive zero-shot generalization capabilities on a wide spectrum of tasks. In this work, we present WeLM: a well-read pre-trained language model for Chinese that is able to seamlessly perform different types of tasks with zero or few-shot demonstrations. WeLM is trained with 10B parameters by "reading" a curated high-quality corpus covering a wide range of topics. We show that WeLM is equipped with broad knowledge on various domains and languages. On 18 monolingual (Chinese) tasks, WeLM can significantly outperform existing pre-trained models with similar sizes and match the performance of models up to 25 times larger. WeLM also exhibits strong capabilities in multi-lingual and code-switching understanding, outperforming existing multilingual language models pre-trained on 30 languages. Furthermore, We collected human-written prompts for a large set of supervised datasets in Chinese and fine-tuned WeLM with multi-prompted training. The resulting model can attain strong generalization on unseen types of tasks and outperform the unsupervised WeLM in zero-shot learning. Finally, we demonstrate that WeLM has basic skills at explaining and calibrating the decisions from itself, which can be promising directions for future research. Our models can be applied from https://welm.weixin.qq.com/docs/api/.
Date	2022-09-21
Short Title	WeLM
Library Catalog	arXiv.org
URL	http://arxiv.org/abs/2209.10372
Accessed	9/27/2022, 7:07:58 AM
Extra	WFZ I like this team's focus on ensuring that the model is "well-read", which is a critical requirement for developing high-quality editorial and publishing

sensibilities. On closer inspection "well read" here means "carefully curated articles from Common Crawl", which is not quite what a book-lover means by being well-read. At the top of their profession authors and editors have read many of the classics; read deeply in at least one modern genre; thought about the trends in the eternal golden dialogue among works; and, most difficult to simulate perhaps, has had a lifetime of charged and transformative encounters with books and the written word. THis is a challenging requirement for pre-training!

DOI	[10.48550/arXiv.2209.10372](https://doi.org/10.48550/arXiv.2209.10372)
Repository	arXiv
Archive ID	arXiv:2209.10372
Date Added	9/27/2022, 7:07:58 AM
Modified	9/27/2022, 7:15:41 AM

- **Tags:**

  - Computer Science - Computation and Language
  - Computer Science - Artificial Intelligence

Attachments

  - arXiv Fulltext PDF
  - arXiv.org Snapshot

- # What Can AI Do to Help the Publishing Industry? Trinka

Item Type	Blog Post
Abstract	Technological developments impact every sector of business, and the publishing industry is no exception. Publishing and media have been undergoing a rapid
Date	2022-02-11
Language	en-US
Short Title	What Can AI Do to Help the Publishing Industry?
URL	https://www.trinka.ai/blog/what-can-ai-do-to-help-the-publishing-industry/
Accessed	7/19/2022, 4:51:25 PM
Extra	WFZ Short survey that accurately describes some existing applications of machine learning in publishing, such as content classification, but discussion of content generation simply restates known current limitations without pushing much further.
Date Added	7/19/2022, 4:51:25 PM
Modified	8/3/2022, 8:04:13 AM

- # Attachments

  - ○ Snapshot

- # What happens when AI has read everything? | LinkedIn

Item Type	Web Page

URL	https://www.linkedin.com/pulse/what-happens-when-ai-has-read-everything-maaike-groenewege/?trackingId=%2Ft6vW0lnAYo48oEbL4EZaQ%3D%3D
Accessed	2/6/2023, 3:41:13 AM
Date Added	2/6/2023, 3:41:13 AM
Modified	2/6/2023, 3:41:13 AM

- # What Is ChatGPT Doing ... and Why Does It Work?

Item Type	Web Page
Abstract	Stephen Wolfram explores the broader picture of what's going on inside ChatGPT and why it produces meaningful text. Discusses models, training neural nets, embeddings, tokens, transformers, language syntax.
Date	2023-02-14
Language	en
URL	https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/
Accessed	4/12/2023, 6:33:24 AM
Date Added	4/12/2023, 6:33:24 AM
Modified	4/12/2023, 6:33:24 AM

- **Attachments**

  - Snapshot

- # What Language Model to Train if You Have One Million GPU Hours?

Item Type	Conference Paper
Author	Teven Le Scao
Author	Thomas Wang
Author	Daniel Hesslow
Author	Lucile Saulnier
Author	Stas Bekman
Author	M. Saiful Bari
Author	Stella Biderman
Author	Hady Elsahar
Author	Jason Phang
Author	Ofir Press
Author	Colin Raffel
Author	Victor Sanh

Author	Sheng Shen
Author	Lintang Sutawika
Author	Jaesung Tae
Author	Zheng Xin Yong
Author	Julien Launay
Author	Iz Beltagy
Abstract	The crystallization of modeling methods around the Transformer architecture has been a boon for practitioners. Simple, well-motivated architectural variations that transfer across tasks and scale...
Date	2022/04/11
Language	en
Library Catalog	openreview.net
URL	https://openreview.net/forum?id=rI7BL3fHIZq
Accessed	6/22/2022, 3:35:50 PM
Conference Name	Challenges {\&
Date Added	6/22/2022, 3:35:50 PM
Modified	12/4/2022, 11:55:24 PM

- **Tags:**

  - #nosource

What Makes Content Engaging? How Emotional Dynamics Shape Success

Item Type	Journal Article
Author	Jonah Berger
Author	Yoon Duk Kim
Author	Robert Meyer
Editor	J. Jeffrey Inman
Editor	Andrew T Stephen
Abstract	Abstract Some cultural products (e.g., books and movies) catch on and become popular, while others fail. Why? While some have argued that success is unpredictable, we suggest that period-to-period shifts in sentiment—what we term sentiment volatility—enhance engagement. Automated sentiment analysis of over 4,000 movies demonstrates that more volatile movies are evaluated more positively. Consistent with the notion that sentiment volatility makes experiences more stimulating, the effect is stronger in genres where evaluations are more likely to be driven stimulation (i.e., thrillers rather than romance). Further, analysis of over 30,000 online articles demonstrate that people are more likely to continue reading more volatile articles. By manipulating sentiment volatility in follow-up experiments, we underscore its causal impact on evaluations, and provide evidence for the role of stimulation in these effects. Taken together, the results shed light on what drives engagement, the time dynamics of sentiment, and cultural analytics or why some cultural items are more successful.

- ## Tags:

  - #nosource

- # What's Really Behind Those AI Art Images?

Item Type	Web Page
Author	Charlie Warzel
URL	https://newsletters.theatlantic.com/galaxy-brain/6317de90bcbd490021b246bf/ai-art-dalle-midjourney-stable-diffusion/
Accessed	9/7/2022, 12:55:47 PM
Extra	WFZ "Thumbsucker" essay, full of thoughtful questions but not getting anywhere really. Contrary to the author I believe the usages are massively transformative and copyright of training data is completely immaterial. As the author notes, the transformation requires extremely complex mathematical processes and compresses down from 100 TB to 2 GB to produce novel high quality works. How exactly can that be considered appropriative? It's like considering all of human culture appropriative.
Website Title	Galaxy Brain
Date Added	9/7/2022, 12:55:47 PM
Modified	9/7/2022, 12:58:57 PM

- # Attachments

  - What's Really Behind Those AI Art Images?

- # What's your type? Try these tests to pick the perfect font for you.

Item Type	Web Page
Author	Emma Kumer
Author	Geoffrey A. Fowler
Author	Leslie Shapiro
Abstract	Fonts can dramatically shape what you communicate and how you read.
Language	en
Short Title	What's your type?
URL	https://www.washingtonpost.com/technology/interactive/2023/best-font-you-test-types-styles/
Accessed	6/27/2023, 6:55:47 PM
Website Title	Washington Post
Date Added	6/27/2023, 6:55:46 PM
Modified	6/27/2023, 6:55:46 PM

- # Attachments

- o Snapshot

When Creation Goes To Zero

Item Type	Web Page
Author	Evan Armstrong
Abstract	Skynet Who?
Date	2022-09-15
URL	https://every.to/napkin-math/when-creation-goes-to-zero
Accessed	3/5/2023, 6:08:37 PM
Date Added	3/5/2023, 6:08:37 PM
Modified	3/5/2023, 6:08:37 PM

- **Attachments**

  - o Snapshot

Why Hollywood Still Goes by the Book

Item Type	Newspaper Article

Author	Francis Levy
Abstract	Article on reliance of moviemakers on books for material for their films; notes movie rights for Michael Crichton novel were recently sold to 20th Century-Fox for large sum, while Peter Benchley was reptdly paid $2.15 million for his next work, The Island; some major novelists are beginning to wield kind of power in Hollywood once held by only 2 or 3 dirs and handful of stars; Mel Sokolow of Casablanca films comments; illus (L)
Date	1979-03-25
Language	en-US
Library Catalog	NYTimes.com
URL	https://www.nytimes.com/1979/03/25/archives/why-hollywood-still-goes-by-the-book-why-hollywood-still-goes-by.html
Accessed	7/25/2023, 8:55:56 PM
Extra	WFZ Recently reupped by the Times as timeless truth.
Section	Archives
Publication	The New York Times
ISSN	0362-4331
Date Added	7/25/2023, 8:55:56 PM
Modified	7/27/2023, 2:22:36 AM

- # Tags:

- o United States
- o Books and Literature
- o Book Trade
- o MOTION PICTURES

Notes:

- o The second and most important reason for the popularity of novels as sources for films is that they frequently break new ground. "Screenwriters tend to be overly aware of what is going on in the industry," says Nancy Hardin, vice president in charge of film and publishing development at Paramount, "whereas a novelist who is writing in an independent environment might come up with something fresh."

Attachments

- o Snapshot

Why is ChatGPT so good? | Blog | Scale AI

Item Type	Web Page
Abstract	How Large Language Models are Trained and Tuned using Reinforcement Learning with Human Feedback (RLHF).

Short Title	Why is ChatGPT so good?
URL	https://scale.com/blog/chatgpt-reinforcement-learning
Accessed	1/26/2023, 2:12:35 PM
Website Title	ScaleAI
Date Added	1/26/2023, 2:12:35 PM
Modified	1/26/2023, 2:12:35 PM

- ## **Attachments**

  - ○ Snapshot

- # Why transformative AI is really, really hard to achieve

Item Type	Web Page
Abstract	A collection of the best technical, social, and economic arguments
Date	2023-06-27T13:38:00+01:00
URL	https://zhengdongwang.com/2023/06/27/why-transformative-ai-is-really-really-hard-to-achieve.html
Accessed	7/5/2023, 1:29:12 PM

- # **Notes:**

  - Aghion, Jones, and Jones explain that the production of ideas itself has steps which are vulnerable to bottlenecks.4 Automating most tasks has very different effects on growth than automating all tasks: ...economic growth may be constrained not by what we do well but rather by what is essential and yet hard to improve... When applied to a model in which AI automates the production of ideas, these same considerations can prevent explosive growth. Consider a two-step innovation process that consists of summarizing papers on arXiv and pipetting fluids into test tubes. Each step depends on the other. Even if AI automates summarizing papers, humans would still have to pipette fluids to write the next paper. (And in the real world, we would also need to wait for the IRB to approve our grants.) In "What if we could automate invention," Matt Clancy provides a final dose of intuition: Invention has started to resemble a class project where each student is responsible for a different part of the

project and the teacher won't let anyone leave until everyone is done... if we cannot automate everything, then the results are quite different. We don't get acceleration at merely a slower rate—we get no acceleration at all. Our point is that the idea of bottlenecking—featured everywhere from Baumol in the sixties to Matt Clancy today—deserves more airtime.5

Attachments

- ○ Snapshot

Workshop on Visualization for AI Explainability

Item Type	Web Page
Abstract	The role of visualization in artificial intelligence (AI) gained significant attention in recent years. With the growing complexity of AI models, the critical need for understanding their inner-workings has increased. Visualization is potentially a powerful technique to fill such a critical need. The goal of this workshop is to initiate a call for 'explainables' / 'explorables' that explain how AI techniques work using visualization. We believe the VIS community can leverage their expertise in creating visual narratives to bring new insight into the often obfuscated complexity of AI systems.
Language	en
URL	http://visxai.io
Accessed	11/2/2022, 8:15:16 PM

Date Added	11/2/2022, 8:15:16 PM
Modified	11/2/2022, 8:15:16 PM

- **Attachments**

  ○ Snapshot

- # WPP teams up with Nvidia to use generative AI in advertising

Item Type	Newspaper Article
Author	Daniel Thomas
Date	2023-05-29
Extra	WFZ I see Don Draper doing the pitch. Free with login/no $.
Publication	Financial Times
Date Added	5/29/2023, 1:30:52 PM
Modified	5/29/2023, 1:31:56 PM

- # Wrestlers « Wrestlers Database « CAGEMATCH - The Internet Wrestling Database

Item Type	Web Page
Author	Philip Kreikenbohm ("Matt Macks")
Date	2001 - present
URL	https://www.cagematch.net/?id=2&view=workers
Accessed	8/21/2022, 7:52:23 PM
Extra	WFZ A robust database of information about a form of scripted narrative entertainment. As a point of reference, the WWE Universe contains 25,559 named characters and 58,545 matches, while the Marvel Universe has ~1200 characters and 40,000 episodes. What I take away from these numbers is that creating artificial narrative universes should be eminently computable.
Date Added	8/21/2022, 7:52:23 PM
Modified	8/21/2022, 8:50:35 PM

- ## Tags:

    - Character network
    - Character creation

Attachments

    - Wrestlers « Wrestlers Database « CAGEMATCH - The Internet Wrestling Database

- # YAKE! Keyword extraction from single documents using multiple local features

Item Type	Journal Article
Author	Ricardo Campos
Author	Vítor Mangaravite
Author	Arian Pasquali
Author	Alípio Jorge
Author	Célia Nunes
Author	Adam Jatowt
Abstract	As the amount of generated information grows, reading and summarizing texts of large collections turns into a challenging task. Many documents do not come with descriptive terms, thus requiring humans to generate keywords on-the-fly. The need to automate this kind of task demands the development of keyword extraction systems with the ability to automatically identify keywords within the text. One approach is to resort to machine-learning algorithms. These, however, depend on large annotated text corpora, which are not always available. An alternative solution is to consider an unsupervised approach. In this article, we describe YAKE!, a light-weight unsupervised automatic keyword extraction method which rests on statistical text features extracted from single documents to select the most relevant keywords of a text. Our system does not need to be trained on a particular set of documents, nor does it depend on dictionaries, external corpora, text size, language, or domain. To demonstrate the merits and significance of YAKE!, we compare it against ten state-of-the-art unsupervised approaches and one supervised method. Experimental results carried out on top of twenty datasets show that YAKE! significantly outperforms other unsupervised methods on texts of different sizes, languages, and domains.

- **Tags:**

  - #nosource
  - Information extraction
  - Keyword extraction

- # Yes, You Can Make an App Too: A Systematic Study of Prompt Engineering in the Automatic Generation of Mobile Applications from User Queries

Item Type	Journal Article
Author	Jasmine Shone
Abstract	App creation should be accessible to everyone, regardless of technical expertise and amount of time. The Aptly platform aims to make this ideal a reality by enabling nearinstantaneous generation of full apps given only an user description of an app idea. In order to optimize Aptly's abilities and tackle a broader question within the large language model community about how to best construct inputs (prompts), we examine the effects of token length, mechanism of choosing examples, and how examples are ordered within the prompt on the quality of Codex's outputs of app code through the construction of a three step automated prompt creation pipeline. We find that these three characteristics affect the quality of the code produced by Codex. Further research is needed on prompt engineering for code-generating large language models such as Codex.
Language	en
Library Catalog	Zotero
Pages	21
Date Added	10/28/2022, 4:26:27 PM

Modified	10/28/2022, 4:26:29 PM

- ## Tags:

- ⊖ No DOI found

Attachments

- Shone - Yes, You Can Make an App Too A Systematic Study o.pdf

---

- # Your future with the Google destroyer | LinkedIn

Item Type	Web Page
Author	Simon Dudley
URL	https://www.linkedin.com/pulse/i-think-ive-just-uncovered-google-killer-simon-dudley/
Accessed	12/6/2022, 4:51:23 PM
Date Added	12/6/2022, 4:51:23 PM
Modified	12/6/2022, 10:26:12 PM

- ## Attachments

•

Item Type	Case
Author	Judge Leonard P. Stark
Short Title	Thaler v. Vidal
Extra	WFZ patent case holds that inventor must be a real, natural person per language of Patent Act. Look to language of Copyright Act for similar answer.
Case Name	STEPHEN THALER v. KATHERINE K. VIDAL, UNDER SECRETARY OF COMMERCE FOR INTELLECTUAL PROPERTY AND DIRECTOR OF THE UNITED STATES PATENT AND TRADEMARK OFFICE
Court	United States Court of Appeals for the Federal Circuit
Date Decided	2022-08-05
Date Added	8/7/2022, 11:45:16 PM
Modified	8/7/2022, 11:53:36 PM