AI & PUBLISHING

AI AND PUBLISHING

Peter Schoppert
NUS Press / Singapore Book Publishers Association
@katong
https://aicopyright.substack.com

PUBLISHERS HAVE A STAKE...

While the cliché is that Large Language Models have been trained by “reading” “the internet”, computer programmes copy, they don’t read. To argue otherwise is to deeply devalue the human act of reading. And the internet is just a channel, publisher & creator content, news, books and journals, makes up a very high proportion of what they models have been trained on.

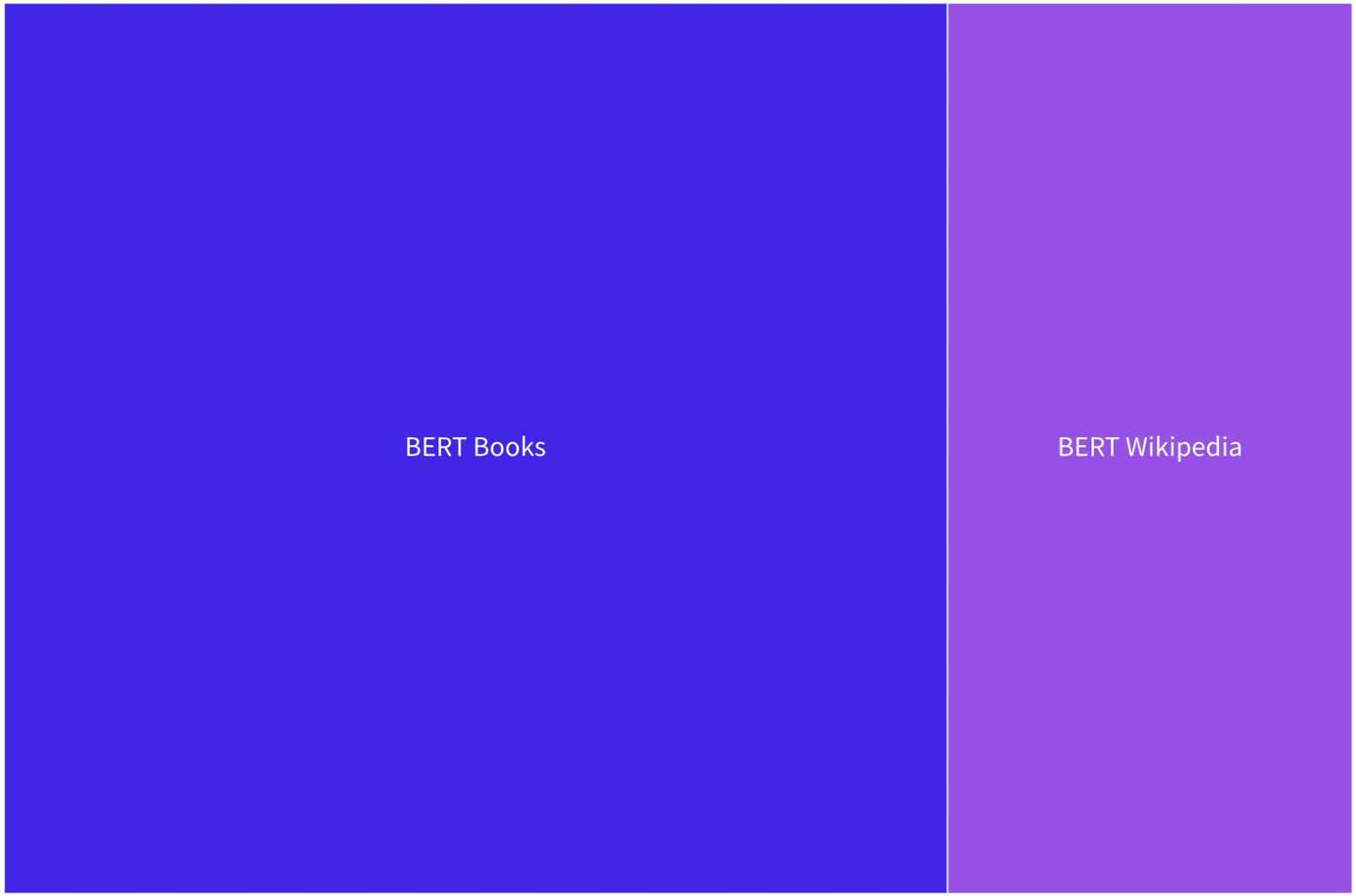
TWO POINTS

- What do we know about training data? And what would it take to be transparent about it?
- The next level of content use in LLM-based systems and how that may be a great opportunity. But can it excuse the original sin?

2 of 5

The first transformer model - trained on books!

The first LLM to use Transformers was BERT from Google, making its debut in 2017. It used two bodies of text for its training: 3b tokens from English Wikipedia, and 7b tokens from BooksCorpus, a dataset based on 11,000 novels downloaded from the Smashwords webset. (Lots of vampires and werewolves!) Bookscorpus had been used since 2015 or so for pre-Transformer models. Total training text size was 10b tokens (roughly 7b words).



We have almost no visibility on what was used to train the models

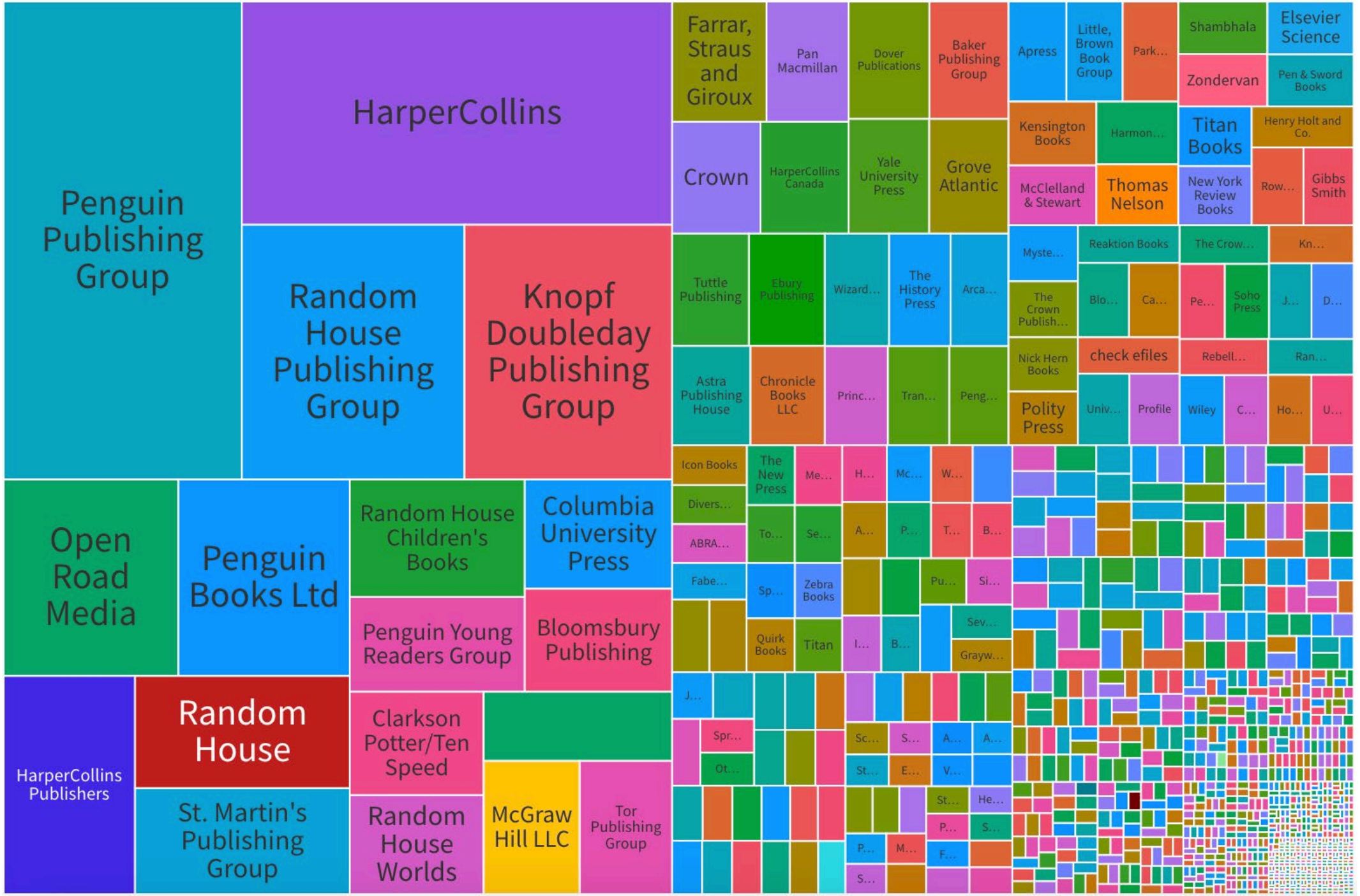
> Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.

Meta's LLaMa-1: 'Unlike Chinchilla, PaLM, or GPT-3, we only use publicly available data, making our work compatible with open-sourcing, while most existing models rely on data which is either not publicly available or undocumented (e.g. "Books – 2TB" or "Social media conversations").'

Meta's LLaMa-2: "a new mix of data from publicly available sources, which does not include data from Meta's products or services. We made an effort to remove data from certain sites known to contain a high volume of personal information about private individuals."

Is it hard to be transparent?

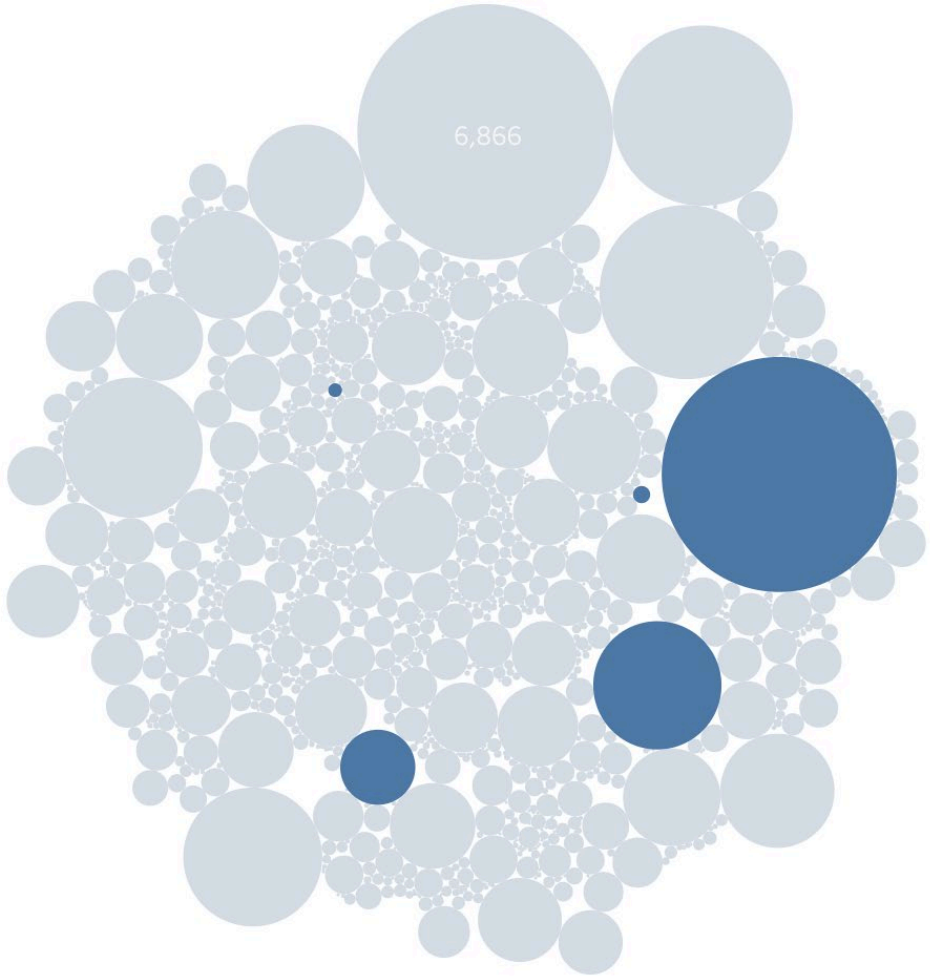
- Books3 has every book in a separate file labelled with its title.
  - I created a searchable list of ISBNs on https://github.com/psmedia/Books3Info
  - *The Atlantic* created a tool to search Books3 by authors
- CommonCrawl webscrapes are lists of URLs
  - The Washington Post has created **"the secret list of websites that make AI like ChatGPT sound smart"** at https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/ - based on Google's C4 dataset.
- LAION (image-text pairs used to train Stable Diffusion) is a list of URLs… and searchable indexes have been created


Penguin Publishing Group
HarperCollins
Random House Publishing Group
Knopf Doubleday Publishing Group
Farrar, Straus and Giroux
Pan Macmillan
Dover Publications
Baker Publishing Group
Apress
Little, Brown Book Group
Park...
Shambhala
Elsevier Science
Zondervan
Pen & Sword Books
Crown
HarperCollins Canada
Yale University Press
Grove Atlantic
Kensington Books
Harmon...
Titan Books
Henry Holt and Co.
McClelland & Stewart
Thomas Nelson
New York Review Books
Row...
Gibbs Smith
Tuttle Publishing
Ebury Publishing
Wizard...
The History Press
Arca...
Myste...
Reaktion Books
The Crow...
Kn...
The Crown Publish...
Blo...
Ca...
Pe...
Soho Press
J...
D...
Nick Hern Books
check efiles
Rebell...
Ran...
Astra Publishing House
Chronicle Books LLC
Princ...
Tran...
Peng...
Polity Press
Univ...
Profile
Wiley
C...
Ho...
U...
Open Road Media
Penguin Books Ltd
Random House Children's Books
Columbia University Press
Icon Books
The New Press
Me...
H...
Mc...
W...
Divers...
To...
Se...
A...
P...
T...
B...
ABRA...
Fabe...
Sp...
Zebra Books
Pu...
Si...
Quirk Books
Titan
I...
B...
Sev...
Grayw...
Penguin Young Readers Group
Bloomsbury Publishing
HarperCollins Publishers
Random House
Clarkson Potter/Ten Speed
St. Martin's Publishing Group
Random House Worlds
McGraw Hill LLC
Tor Publishing Group
J...
Spr...
Ot...
Sc...
S...
A...
St...
E...
V...
St...
He...
P...
S...
P...
M...
F...
S...

ISBNS from the Books3 database by Peter Schoppert

publishers with more than 350 I... | By Publisher (bubble) | By Imprint (bubble)

Publisher by number of ebook ISBNs in Books3 dataset.



Highlight Publisher

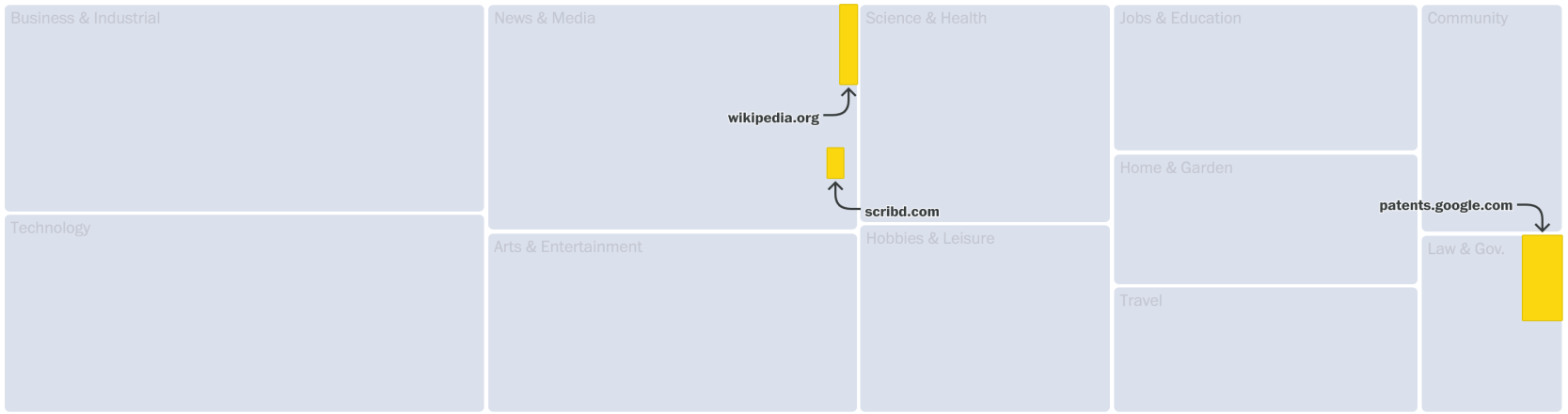
Harper

HarperChristian Resources
HarperCollins
HarperCollins Canada
HarperCollins Leadership
HarperCollins Publishers

tableau

Inside the secret list of websites that make AI like ChatGPT sound smart

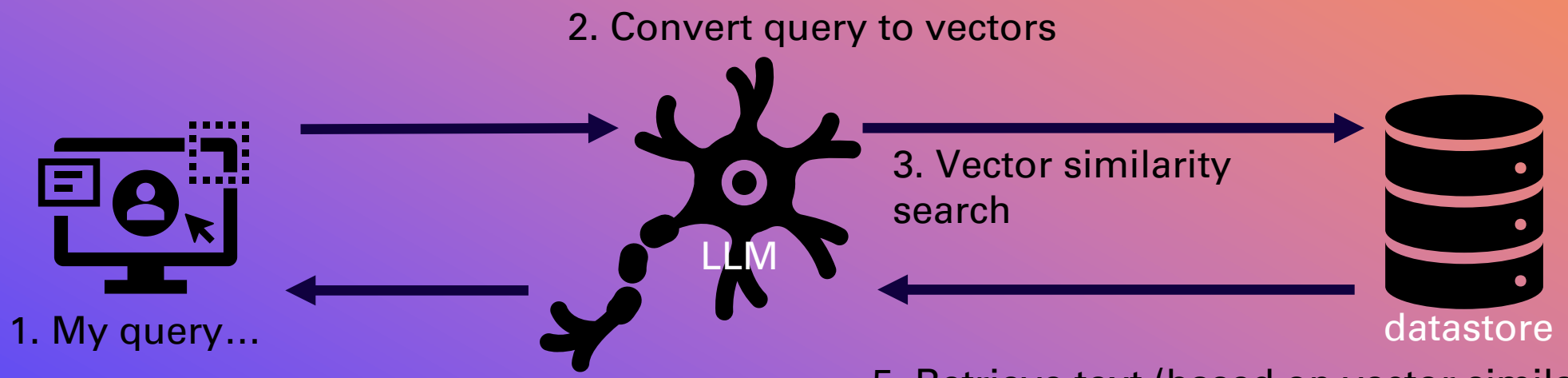
Washington Post, April 19, 2023



Wikipedia to Wowhead

The data set was dominated by websites from industries including journalism, entertainment, software development, medicine and content creation, helping to explain why these fields may be threatened by the new wave of artificial intelligence. The three biggest sites were patents.google.com No. 1, which contains text from patents issued around the world; wikipedia.org No. 2, the free online encyclopedia; and scribd.com No. 3, a subscription-only digital library. Also high on the list: b-ok.org No. 190, a notorious market for pirated e-books that has since been seized by the U.S. Justice Department. At least 27 other sites identified by the U.S. government as markets for piracy and counterfeits were present in the data set.

RETRIEVAL AUGMENTED GENERATION



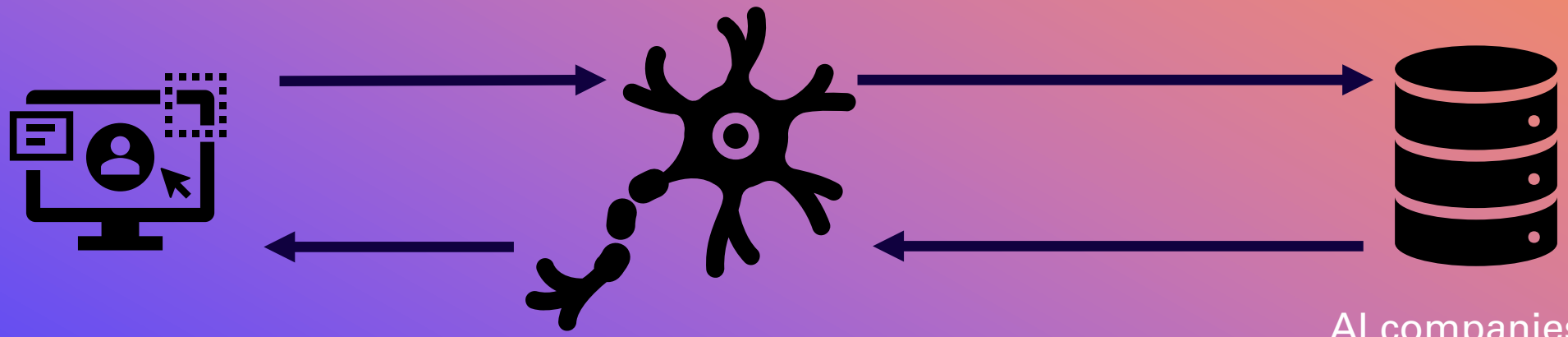
5. Retrieve text (based on vector similarity) and send it back to LLM in the form of a prompt:

```prompt-context
Anticipate the type of answer desired by the user. Imagine the following 3 articles were written by the user and contain all the necessary information to answer the user's question. Begin responses with ”According to Publisher X data...” and include citations
---BEGIN #1---
---END #1---
```

6. LLM processes prompt with context and replies to query

“According to articles from Publisher X…

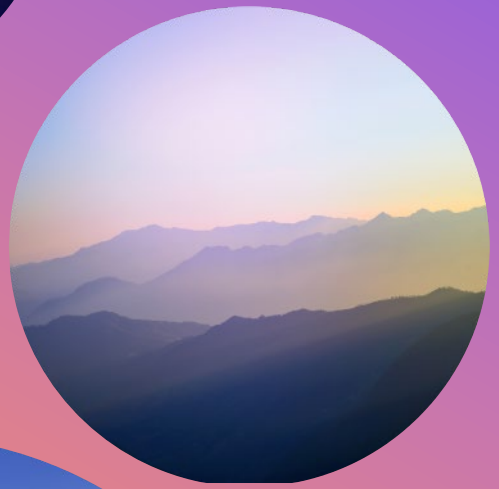
LICENSING



They are still refusing to accept that they should have licensed the data used to train the base model...

AI companies are now offering to license this material. This is how Chat-GPT can be up-to-date by licensing news data. Each query can be traced to specific unit of text.

In European case, “opting out” is the best first step to being able to license.



THANK YOU

Peter Schoppert
Schoppert@nus.edu.sg
https://aicopyright.substack.com
https://psmedia.asia