

The Predominant Use of High-Authority Commercial Web Publisher Content to Train Leading LLMs

George Wukoson, Lead Attorney on AI Matters, and
Joey Fortuna, Chief Technology Officer
Ziff Davis

I. Introduction

Large language models predict text from preceding text. Developers train LLMs by ingesting text data.¹ LLMs' language faculty has emerged from the vast scale of their training corpora. LLM developers achieve that scale by copying and ingesting vast quantities of text from the web.

But not all web text is equally valuable to LLM developers.

In this paper, we review published details of training data from major LLM company research teams and analyze the datasets. Our work shows that key LLM training datasets are disproportionately composed of high-quality content owned by commercial publishers of news and media websites. Major LLM companies have quantifiably prioritized this content in training the most important LLMs over the short history of the technology.²

As LLMs have evolved from pure research projects to some of the most valuable IP assets on earth, LLM companies have ceased publishing training details, and publishers have brought litigation against them. Courts and policymakers are grappling with questions of IP rights and technological progress.³ We attempt to illuminate the richest available sources of information on LLM companies' use of web publisher content to inform this vital public conversation.

¹ In this paper, we generally use "trained" synonymously with "pre-trained," though development stages differ by model.

² Our analysis follows on and is educated by the work of Prof. Vince Conitzer, PhD, Aakar Gupta, PhD, Eric Sodomka, PhD, and their team. "White Paper: How the pervasive copying of expressive works to train and fuel generative artificial intelligence systems is copyright infringement and not a fair use," <https://www.newsmediaalliance.org/wp-content/uploads/2023/10/AI-White-Paper-with-Technical-Analysis.pdf> (2023). Unlike the Conitzer group, we focused on documents (identified by URLs) rather than tokens, for the sake of workability when dealing with datasets as large as Common Crawl and when analyzing hundreds of domains rather than a discrete set (16 in the case of Conitzer, et al.). We also draw on insights from the Allen Institute for AI's and The Washington Post's analysis of the C4 dataset. Dodge, et al., "Documenting Large Webtest Corpora: A Case Study on the Colossal Clean Crawled Corpus," <https://arxiv.org/pdf/2104.08758> (2021); Schaul, et al., "Inside the secret list of websites that make AI like ChatGPT sound smart," <https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/> (April 19, 2023).

³ In addition to litigation discussed in this paper, for recent relevant legislative proposals, see the AI Foundation Model Transparency Act (H.R. 6881; <https://www.congress.gov/bill/118th-congress/house-bill/6881/text>), which would direct federal agencies to establish rules for model developer disclosure of, among other things, training data, and the Generative AI Copyright Disclosure Act (H.R. 7913; <https://www.congress.gov/bill/118th-congress/house-bill/7913/text>), which would require training dataset creators to disclose a detailed summary of any copyrighted works used to the U.S. Copyright Office.

II. Datasets

We analyzed four datasets: Common Crawl, the Colossal Clean Crawled Corpus (C4), OpenWebText, and OpenWebText2. Each of these datasets is a corpus of text scraped from the web and processed according to a published recipe. They are distinguished from one another by the nature of their processing—from mere text extraction, to simple cleaning, to extensive curation—and resulting level of abstraction from the raw web. We analyzed these datasets because they or the proprietary datasets they represent have been used to train enterprise-making breakthrough commercial LLMs. We see evidence that they continue to underlie or influence current frontier models.

A. Common Crawl

*The copy of Common Crawl we analyzed comprises 3.15 billion web pages.*⁴

Common Crawl⁵ is a nonprofit organization that maintains an open repository of web crawl data. It crawls the entire web about once per month and releases datasets from those crawls. These datasets are used for LLM training and other purposes. Common Crawl predates LLMs.

Each release of Common Crawl (or CC) contains full raw web page data and separate extracts of metadata and text. We focused on the text extract set because it is used for model training and serves as a whole-web control to test the effect of cleaning and curation on the makeup by source of purpose-built LLM text corpora.

B. C4

The copy of C4 we analyzed comprises 365 million web pages.

The Colossal Clean Crawled Corpus, or C4, is a cleaned-up Common Crawl.⁶ It was created impartially, in that it is cleaned but not curated or otherwise optimized for quality. To create C4, a CC snapshot was filtered to remove non-English pages, "gibberish" text, menus, boilerplate, code, very abbreviated or offensive text, and duplicates.⁷ This methodology was inspired by earlier efforts to clean up CC data for natural language processing work.

Google created C4 and used it to train its T5 and LaMDA models. Meta also used C4 to train its LLaMA model.

C. OpenWebText

The copy of OpenWebText we analyzed comprises 23 million web pages.

OpenWebText is an open-source proxy for OpenAI's WebText dataset, created to train GPT-2.

⁴ <https://commoncrawl.org/blog/july-august-2021-crawl-archive-available>.

⁵ <https://commoncrawl.org/>.

⁶ We refer to English C4, or C4/en, as C4. It was the first among multiple variants, described here by configuration: <https://www.tensorflow.org/datasets/catalog/c4>.

⁷ Further described by Raffel, et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," <https://arxiv.org/pdf/1910.10683>, p. 6, *Journal of Machine Learning Research* 21 (2020).

An OpenAI research team created WebText to “emphasize[] document quality” by “only scrap[ing] web pages which have been curated/filtered by humans.”⁸ Given the cost of manual curation at scale, they crowdsourced curation to users of Reddit, scraping the web pages at all outbound links from Reddit posts and comments with at least three “karma” (upvotes) as a signal reflecting human interest, educational value, or humor. This yielded 45 million web pages. The OpenAI team extracted the text from the HTML, deduplicated the documents, cleaned the text, and removed Wikipedia pages, which would overlap with another GPT-2 training corpus.

The OpenAI team disclosed the steps it took to create WebText in its GPT-2 paper⁹ and released a list of the top 1,000 domain names in WebText by document frequency¹⁰ as well as WebText data samples¹¹. The full WebText dataset was not made publicly available. Two scholars set out to replicate WebText. They created OpenWebText by following the steps OpenAI disclosed—acknowledging that the reproduction would not be perfect.¹² They released OpenWebText effectively into the public domain, designating it CC0.¹³

While reproducing a dataset from a few paragraphs in a paper is inherently imperfect, OpenAI’s recipe was a simple one to follow: scrape all of the pages linked to in Reddit posts and comments with at least three karma and perform the same simple cleaning that the OpenAI team did. As a result, the free and public OpenWebText substantially overlaps with known contents of WebText and is an effective proxy for it.

D. OpenWebText2

The copy of OpenWebText2 we analyzed comprises 17 million web pages.

OpenWebText2 is an open-source proxy for OpenAI’s WebText2 dataset, which it created to train GPT-3.

An OpenAI research team described WebText2 in its July 2020 GPT-3 paper as an expanded version of WebText, created by scraping Reddit link pages over a longer period of time and totaling 19 billion tokens in size.¹⁴ The initial version of GPT-3 was trained on a dataset created in November 2016. In September 2020, an updated version of GPT-3 was released reflecting web content up to August 2020.¹⁵

⁸ Radford, et al., “Language Models are Unsupervised Multitask Learners,” https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf, p.3 (2019).

⁹ Ibid.

¹⁰ gpt-2/domains.txt, <https://github.com/openai/gpt-2/blob/master/domains.txt> (as referenced in GPT-2 model card, https://github.com/openai/gpt-2/blob/master/model_card.md).

¹¹ gpt-2-output-dataset, <https://github.com/openai/gpt-2-output-dataset>.

¹² OpenWebTextCorpus, <https://skylion007.github.io/OpenWebTextCorpus/>.

¹³ CC0 is a legal mechanism developed by Creative Commons to waive copyright interests to the maximum extent possible, a difficult task under copyright law alone. See <https://creativecommons.org/public-domain/cc0/>.

¹⁴ Brown, et al., “Language Models are Few-Shot Learners,” <https://arxiv.org/pdf/2005.14165> (2020). A token is a unit of text processed by a language model, which might range in length from a single character to an entire word. There are various methods of tokenizing text that will yield different and differently-sized tokens.

¹⁵ GPT-3 Model Card, <https://github.com/openai/gpt-3/blob/master/model-card.md>.

EleutherAI, a non-profit AI research group, created OpenWebText2 as a follow-up to OpenWebText by scraping web pages at Reddit links with at least three karma over an extended and updated period of time—from 2005 through April 2020, near GPT-3’s training data cutoff.¹⁶ EleutherAI released OpenWebText2 under the open-source MIT license.

OpenAI did not release a step-by-step “recipe” for WebText2 like it did for WebText. It more generally described it. OpenWebText2’s precision as a copy is therefore hazier than its predecessor. However, WebText2 and OpenWebText2 each are built from a simple common concept, though not a recipe: scrape Reddit links with at least 3 karma over long periods of time to create large datasets. They should therefore bear a directional resemblance to one another, and OpenWebText2 is at least a presumptive proxy for the proprietary WebText2.

E. Use in Breakthrough Model Training

CC, C4, WebText, and WebText2 were used to train the breakthrough LLMs that laid the foundation for today’s leading commercial LLM businesses. Here are the LLMs that were trained on these datasets, along with their other known training data sources:

	GPT-2 ¹⁷	GPT-3 ¹⁸	T5 ¹⁹	LaMDA ²⁰	LLaMA ²¹
Common Crawl		✓ (filtered)			✓ (filtered)
C4			✓	✓	✓
Other web				✓	✓
“Public forums”				✓	
WebText	✓				
WebText2		✓			
Wikipedia		✓	✓	✓	✓
Books corpora		✓			✓

Notably, WebText2 was given 22% weight in training GPT-3 despite accounting for 3.8% of tokens, per OpenAI’s disclosure in its GPT-3 paper:

¹⁶ Gao, et al., “The Pile: An 800GB Dataset of Diverse Text for Language Modeling,” <https://arxiv.org/pdf/2101.00027>, p. 23 (2020); <https://openwebtext2.readthedocs.io/>.

¹⁷ GPT-2 model card, https://github.com/openai/gpt-2/blob/master/model_card.md.

¹⁸ Brown, et al., pp. 9, 43; GPT-3 Model Card.

¹⁹ Model Card for T5 Base, <https://huggingface.co/google-t5/t5-base#training-details>.

²⁰ Thoppilan, et al., “LaMDA: Language Models for Dialog Applications,” <https://arxiv.org/pdf/2201.08239> (2022).

²¹ Touvron, et al., “LLaMA: Open and Efficient Foundation Language Models,” <https://arxiv.org/pdf/2302.13971> (2023).

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Source: Brown, et al., p. 9

In that paper, the OpenAI authors additionally stated: “during training, datasets are not sampled in proportion to their size, but rather datasets we view as higher quality are sampled more frequently, such that CommonCrawl and Books2 datasets are sampled less than once during training, but the other datasets are sampled 2-3 times,” with WebText2 being sampled 2.9 times (“epochs”) as shown above.²² OpenAI’s oversampling and overweighing of WebText2 shows that it prioritized high-quality web text datasets at a key moment in its enterprise and product development history. As we will see, the high quality of WebText2 is driven by its disproportionate composition of commercially produced web media content.

Meta found another way of emphasizing curated web text.

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

Source: Touvron, et al., p. 2

While on its face simply sampling Common Crawl and C4 without disproportionate weight or frequent sampling (figure from LLaMA paper above), Meta crowdsourced curation of Common Crawl to the human writers of Wikipedia. It first used an existing algorithm to deduplicate the CC data, remove non-English pages, and remove low-quality content. It then trained and used a text classification model to find pages in the filtered CC set that are used as reference citations in human-written Wikipedia articles and discard those that are not.²³ The resulting “Common Crawl” was thus both filtered and curated for quality in a manner similar to WebText and WebText2 (but outsourcing via Wikipedia rather than Reddit).

In addition to serving as proxies for WebText and WebText2, OpenWebText and OpenWebText2 have been used by major LLM developers to train research and commercial models. Anthropic trained an LLM on OpenWebText itself.²⁴ And Apple, Nvidia, and Salesforce all used

²² Brown, et al., pp. 8–9. An epoch is a complete pass through a corpus in training.

²³ Touvron, et al., p. 2.

²⁴ Askell, et al., “A General Language Assistant as a Laboratory for Alignment,” <https://arxiv.org/pdf/2112.00861>, 27 (2021).

OpenWebText2 itself—as part of a compilation of training datasets called the Pile, also released by EleutherAI²⁵—to train LLMs:

- Apple trained its LLM OpenELM on the Pile, among other datasets.²⁶
- Nvidia trained its LLM NeMo Megatron-GPT 20B on the Pile.²⁷
- Salesforce trained its LLM CodeGen on the Pile, among three datasets.²⁸
- BloombergGPT trained BloombergGPT on the Pile (which it notes contains OpenWebText2), C4, and a proprietary web text dataset that makes up the bulk of Bloomberg's proprietary FinPile dataset.²⁹
- Databricks trained its LLM Dolly v2 on the Pile.³⁰

F. Relation to Frontier Models

These foundational models are the progenitors of today's frontier models. While those frontier models' training is kept secret, we find evidence that the older public training sets still influence the new models.

i. OpenAI

OpenAI's GPT-2 and GPT-3 directly preceded GPT-3.5, the original model behind ChatGPT and a key catalyst of the soaring rise of generative AI and LLM company valuations with it. GPT-4 and GPT-4o followed GPT-3/3.5. OpenAI has declined to disclose the training data for GPT-4 and GPT-4o, as discussed below in **Section II.F.iv**.

GPT-4o is now considered the foremost LLM. OpenAI was most recently valued at \$80 billion or more.³¹

ii. Google

Google's LaMDA was the model underlying Google's initial release of its Bard chatbot and its Search Generative Experience LLM search engine product.³² The LaMDA family of models preceded Google's PaLM family of models, and its current frontier model family Gemini followed after that.

²⁵ Described in Gao, et al.

²⁶ Mehta, et al., "OpenELM: An Efficient Language Model Fram

²⁷ <https://huggingface.co/nvidia/nemo-megatron-gpt-20B#training-data>.

²⁸ Nijkamp, et al., "CodeGen: An Open Large Language Model for Code with Multi-Turn Program Synthesis," <https://arxiv.org/pdf/2203.13474>, p. 3 (2023).

²⁹ Wu, et al., "BloombergGPT: A Large Language Model for Finance," <https://arxiv.org/pdf/2303.17564>, pp. 5–9 (2023).

³⁰ <https://huggingface.co/databricks/dolly-v2-3b>; <https://huggingface.co/databricks/dolly-v2-7b>;

³¹ "OpenAI Completes Deal That Values the Company at \$80 Billion," *The New York Times*, <https://www.nytimes.com/2024/02/16/technology/openai-artificial-intelligence-deal-valuation.html> (Feb. 16, 2024).

³² Sundar Pichai, "An important next step on our AI journey," The Keyword, <https://blog.google/technology/ai/bard-google-ai-search-updates/> (Feb. 6, 2023).

Alphabet's market capitalization is currently \$2.0 trillion. CEO of Google DeepMind Demis Hassabis recently stated that Google will spend over \$100 billion developing AI technology.³³

iii. Meta

Meta's Llama model family³⁴ has rapidly developed from its first version. Meta released each of three versions of Llama under a custom, nominally open-source license. It no longer discloses Llama training data in detail, but it states that Llama 2 and 3 were trained on "[a] new mix of publicly available online data."³⁵ Meta uses Llama 3 to power a multimodal generative AI chatbot called Meta AI, offered standalone and in Meta's flagship social media platforms Facebook, Instagram, WhatsApp, and Messenger.³⁶

Meta's market capitalization is currently \$1.5 trillion. In its first quarter 2024 earnings press release, Meta stated that it increased its full-year 2024 capital expenditures to \$35–40 billion from \$30–37 billion in connection with AI infrastructure investments and that it anticipates capital expenditures to increase in 2024 "as we invest aggressively to support our ambitious AI research and product development efforts."³⁷

iv. Current Disclosure Trends

Major LLM developers no longer disclose their training data as they once did. They are now more commercial and less transparent. (We will discuss IP litigation; it has surely contributed to their opacity.) OpenAI and Google have not disclosed the training data for their latest frontier models, and neither have Meta, Anthropic, Cohere, or Mistral.

When these companies do comment on training data, they emphasize their use of "publicly available" internet data, omitting the details. For example, OpenAI states that it trained GPT-4 on "both publicly available data (such as internet data) and data licensed from third-party providers."³⁸ Anthropic—which has never published training details—states that "Claude models are trained on a proprietary mix of publicly available information from the Internet, datasets that we license from third party businesses, and data that our users affirmatively share or that crowd workers provide."³⁹ Meta's similar statement is quoted above in **Subsection iii**.

³³ "DeepMind CEO Says Google Will Spend More Than \$100 Billion on AI," *Bloomberg*, <https://www.bloomberg.com/news/articles/2024-04-16/deepmind-ceo-says-google-will-spend-more-than-100-billion-on-ai> (April 15, 2024).

³⁴ Meta ceased styling these models "LLaMA" and began styling them "Llama" beginning with Llama 2.

³⁵ Model Details, https://github.com/meta-llama/llama/blob/main/MODEL_CARD.md. Model Details, https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

³⁶ "Meet Your New Assistant: Meta AI, Built With Llama 3," <https://about.fb.com/news/2024/04/meta-ai-assistant-built-with-llama-3/> (April 18, 2024).

³⁷ Press release, "Meta Reports First Quarter 2024 Results," <https://investor.fb.com/investor-news/press-release-details/2024/Meta-Reports-First-Quarter-2024-Results/default.aspx> (Apr. 24, 2024).

³⁸ OpenAI, "GPT-4 Technical Report," <https://arxiv.org/pdf/2303.08774>, p. 2 (2024). OpenAI elsewhere states, "GPT-4 has learned from a variety of licensed, created, and publicly available data sources...." GPT-4 System Card, <https://cdn.openai.com/papers/gpt-4-system-card.pdf> (2023).

³⁹ Anthropic, "Model Card and Evaluations for Claude Models," <https://www-cdn.anthropic.com/bd2a28d2535bfb0494cc8e2a3bf135d2e7523226/Model-Card-Claude-2.pdf>, p. 2 (2023).

Despite this obfuscation, the current models have more than heritage to tie them to the earlier, disclosed datasets containing high-quality curated web content. That kind of content is apparently necessary to build an LLM. It was indispensable to training GPT-3, LLaMA, and other breakthrough models, and we are not aware of any fundamental architectural change that would cause that to change.⁴⁰ In fact, LLM company statements made over the past year about licensing deals with commercial web publishers indicate that the need for high-quality training text data has only grown more acute as developers compete to keep scaling. For example:

- “[A]ccess to [AP’s] high-quality, factual text archive...will help to improve the capabilities and usefulness of OpenAI’s systems.”
–Brad Lightcap, OpenAI COO, on OpenAI’s content licensing deal with the Associated Press⁴¹
- “The initiative will enrich users’ experience with ChatGPT by adding recent and authoritative content on a wide variety of topics, and explicitly values the publisher’s role in contributing to OpenAI’s products.”
–Joint press release on OpenAI’s content licensing deal with Axel Springer⁴²
- “We’re joining forces with News Corp to...enrich our products with its premium journalism....Through this partnership, OpenAI has permission to...enhance its products, with the ultimate objective of providing people the ability to make informed choices based on reliable information and news sources.”
–Joint press release on OpenAI’s content licensing deal with News Corp⁴³

Additionally, a recent model and AI product are directly tied to one of the datasets we analyzed. Deep academic scrutiny of OpenAI (as the leading LLM company) reveals a throughline from WebText2 to GPT-3.5/ChatGPT. Between the original GPT-3 trained on WebText2 and GPT-3.5 underlying ChatGPT, OpenAI created numerous intervening models without fully disclosing how each relates to the last. It is a complicated family tree. But a detailed analysis by a scholar at the Allen Institute for AI found that the foundational pre-training of the whole GPT-3/3.5 family occurred at the initial GPT-3/WebText2 stage:⁴⁴

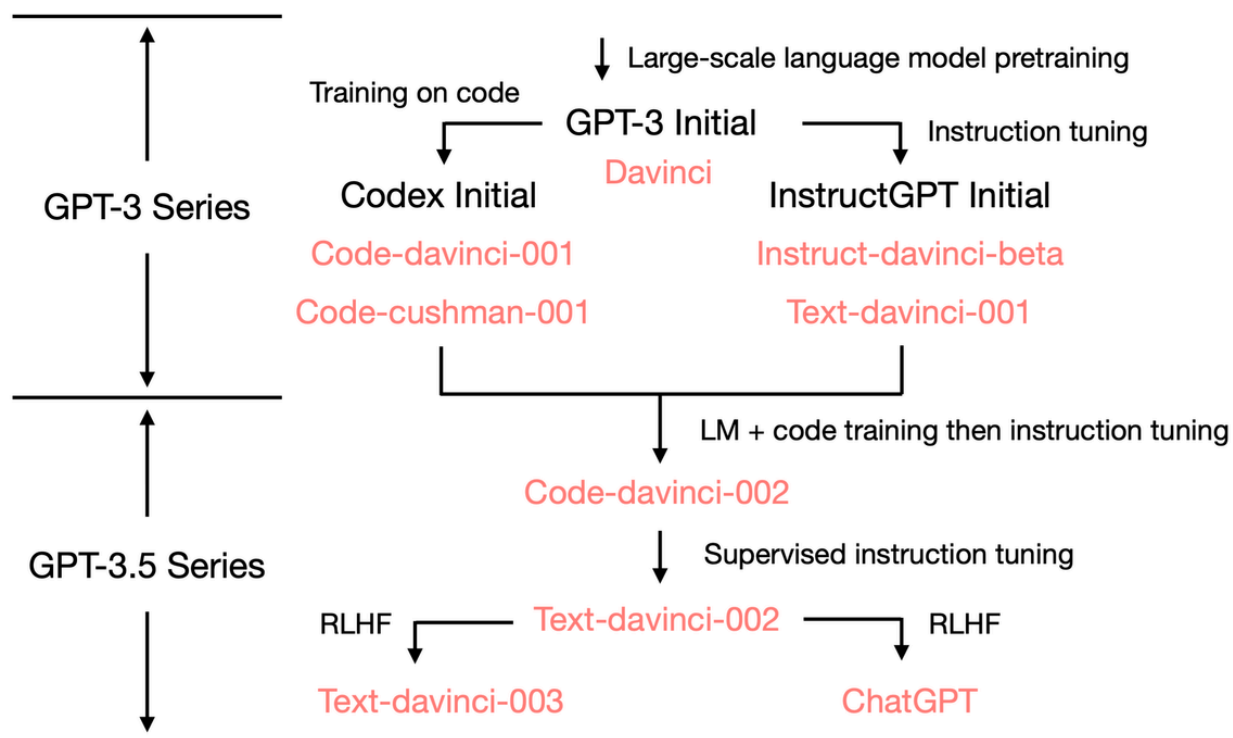
⁴⁰ Many of the frontier models are multimodal, and so have been trained on images and audio in addition to text. But this has not changed the nature of text pre-training.

⁴¹ Press release, “AP, OPEN AI AGREE TO SHARE SELECT NEWS CONTENT AND TECHNOLOGY IN NEW COLLABORATION,” <https://www.ap.org/media-center/press-releases/2023/ap-open-ai-agree-to-share-select-news-content-and-technology-in-new-collaboration/> (July 13, 2023).

⁴² Press release, “Axel Springer and OpenAI partner to deepen beneficial use of AI in journalism,” <https://www.axelspringer.com/en/ax-press-release/axel-springer-and-openai-partner-to-deepen-beneficial-use-of-ai-in-journalism>.

⁴³ Press release, “A landmark multi-year global partnership with News Corp,” <https://openai.com/index/news-corp-and-openai-sign-landmark-multi-year-global-partnership/> (May 22, 2024).

⁴⁴ “How does GPT Obtain its Ability? Tracing Emergent Abilities of Language Models to their Sources,” <https://yaofu.notion.site/How-does-GPT-Obtain-its-Ability-Tracing-Emergent-Abilities-of-Language-Models-to-their-Sources-b9a57ac0fc74f30a1ab9e3e36fa1dc1> (May 16, 2023).



This allows at least a working presumption that the value of that training accrued to the succeeding models.

III. Publishers

We chose a set of leading commercial publishers to query for in the LLM training datasets based on contemporaneous rankings of web publisher aggregate traffic. Specifically, we pulled web traffic rankings (by unique visitors) in Comscore—the leading traffic measurement company for web advertising—for the Media category on desktop and mobile⁴⁵ from August 2020, the last month of data for the most recently updated relevant dataset (WebText2). We discretionarily filtered the rankings for the top 15 portfolio publishers, excluding: single-property publishers, non-media tech firms, other platforms for user-generated content, and entertainment firms (except for News Corp, which has a substantial digital publishing business). Where a publisher in the ranking was a subsidiary of a larger organization, we substituted the larger organization, and in some cases multiple publishers in the relevant rankings tier were thereby consolidated. For example, we substituted News Corp for Fox Corporation and NY Post Network.

Applying this methodology, we arrived at the following set—with relevant subsidiary brands in parentheses: Advance (Condé Nast, Advance Local), Alden Global Capital (Tribune Publishing, MediaNews Group), Axel Springer, Bustle Digital Group, BuzzFeed, Inc., Future plc, Gannett, Hearst, IAC (Dotdash Meredith and other divisions), News Corp, The New York Times Company, Penske Media Corporation, Vox Media, The Washington Post, and Ziff Davis.

⁴⁵ This excludes a third category of traffic that Comscore tracks: social. Social would consist of non-web (i.e., proprietary platform) traffic and would not be relevant to determining publishers to query for in web text datasets.

Compiling the domain names of these publishers also required judgment. We are not aware of any published single source of each of these publishers' domain names. It was not practical to use WHOIS for this purpose for two reasons: (1) following the entry into force of the GDPR, WHOIS records are often anonymized, and (2) to the extent they are not anonymized, these companies' corporate structures would make for numerous different entity registrants per company, making list compilation difficult. So, we instead collected lists of domain names for each of these publishers through manual review, primarily based on the companies' lists of their brands on their industry-facing marketing web pages (such as condenast.com/brands). The publishers' domains are listed in **Appendix 1** to this white paper.

IV. Domain Authority

Domain Authority (or DA) is a widely-used metric originally developed by Moz, a search engine optimization company.⁴⁶ The algorithm used to determine Domain Authority is proprietary. In general terms, the metric is designed to assign a quantitative value to the "rankability" of a particular domain. Domains and webpages are given a score from 1 to 100, with 100 being the most likely to enjoy high ranking on search engine result pages (SERPs). Moz derives this value by incorporating data from its repository of linking root domains and the overall number of links back to the domains represented.⁴⁷ It considers other attributes, such as the age of the domain and the size of the website (in pages published).⁴⁸

Domain Authority is a metric of a website's extrinsic quality as implicitly endorsed by other web publishers—by way of references to, mentions of, and links back to the scored domains. It is not an evaluation of the intrinsic quality of written text. It is a distillation of objective indicators of a site's relevance and relativistic influence. Moz's Domain Authority score is widely regarded as an informative and authoritative measurement of quality in this sense.

Although Moz.com was the first SEO company to develop an index of this nature, comparable metrics have been developed by competitors to Moz, such as Ahref's "Website Authority" (formerly: "Domain Rating"). We carried out a limited comparison of Moz's Domain Authority with Ahref's Website Authority and, across a list of 55 domains, found their score values to be within 1.5 points of each other, with Moz scores being slightly higher overall. The absolute values of the scores differed by an average of six points. When grouped by score range, we found the differences at the high end (in the 80–100 range) to be within 1.6 points in actual value and within 2 points in absolute value, with Moz again being higher. At the mid-range (50–60), we found the absolute difference to be around 5.5 and actual differences at around -1.4, with Ahrefs trending higher. And at the low range (0–30), the absolute difference was around 8 points and the difference in actual value closer to 4, with Moz trending higher (see Appendix 2).

Separate analysis could be carried out weighing each of these metrics against the others, but as they are all relativistic quantitative approaches (meaning: each domain is assigned a score

⁴⁶ Moz is a portfolio company of Ziff Davis, the authors' employer.

⁴⁷ Moz, "Authority Scoring: What It Is, How It's Changing, and How to Use It," https://moz-static.s3.amazonaws.com/products/landing-pages/announcements/Authority_Scoring_Guide.pdf.

⁴⁸ The HOTH, "Moz Domain Authority Vs. Ahrefs Domain Rating: Decoding the Metrics," <https://www.thehoth.com/blog/da-vs-dr/> (Mar. 30, 2023).

relative to all other domains in the auditor’s database), and each approach is rooted in the domain’s influence as measured by the number of backlinks aiming at that domain, the overall distribution of authority should be hold regardless of the method of evaluation.

In choosing among SEO data sources (including Moz and Ahrefs), database domain composition might cause minor variations in Domain Authority/Website Authority-based analyses. (Moz’s database of domains may contain domains not found in Ahref’s database and vice versa.) However, given the volume of domains under analysis, our view is that such variations would not lead to a material difference in findings.

V. Analysis

A. Overview

The corpora we analyzed range from an uncurated blind crawl of the entire internet (Common Crawl) to a rigorously curated dataset (OpenWebText2), produced through methods of cleaning and curating employed described in **Section II** above.

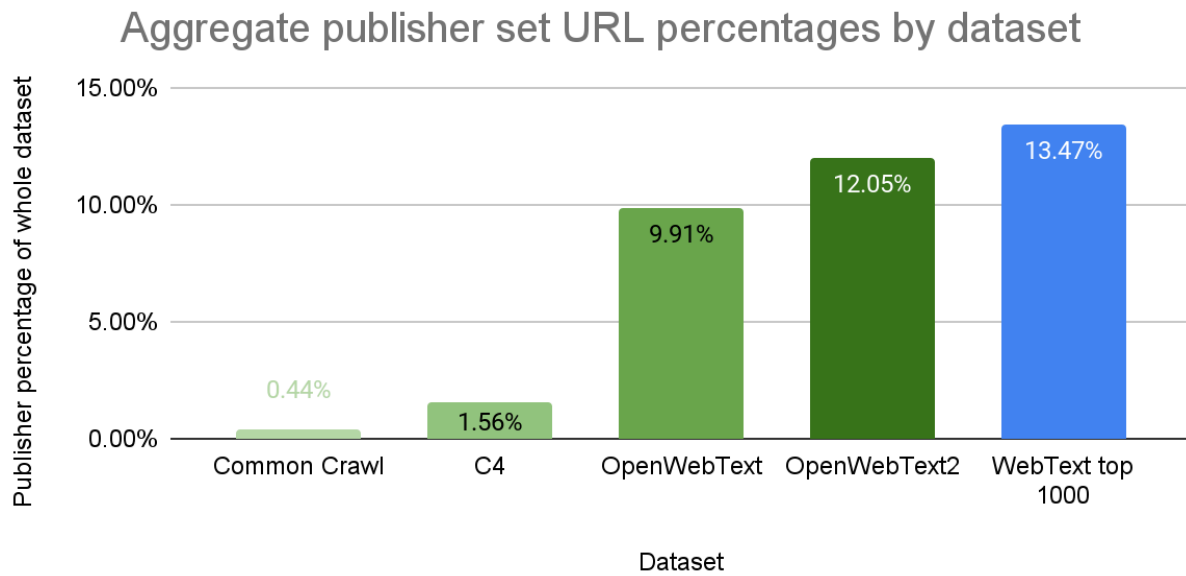
We show two positive correlations flowing from curation:

First, increasing curation is correlated with more commercial publisher sourcing. As we move along the spectrum from uncurated to highly-curated datasets, news and media websites owned by the commercial publishers become more and more concentrated.

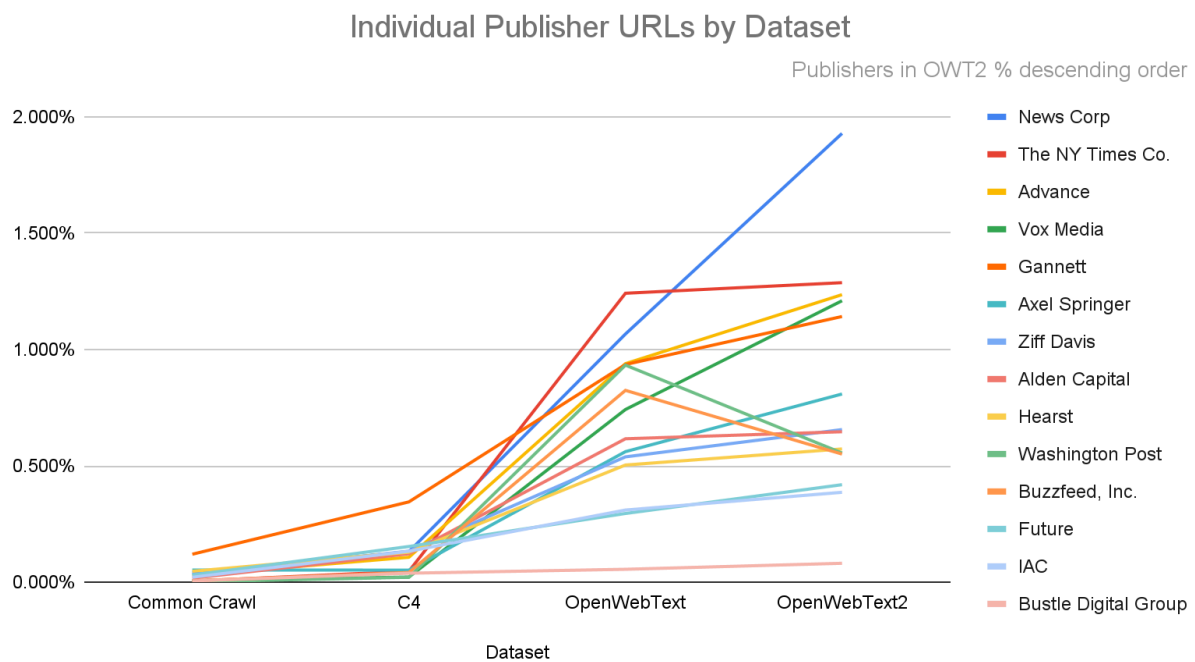
Second, increasing curation is correlated with higher Domain Authority. URLs from high DA sites appear in greater proportion in more curated corpora, and low-DA sites predominate in the uncurated corpora.

B. Commercial Web Publisher Share of Dataset

When we aggregated counts of web documents by URL in the public LLM datasets we analyzed, we saw that the key publishers we selected made up a small percentage of uncurated Common Crawl—only 0.44%. The publishers’ proportion tripled to 1.55% in cleaned (but not curated for quality) C4. Presence percentage increased precipitously in OpenWebText and OpenWebText2, more than sextupling to 9.91% in OpenWebText and rising further to 12.04% in OpenWebText2. Finally, the publishers had a 13.47% share of URLs in the top 1,000 domains in WebText, as directly disclosed by OpenAI.



This relationship also generally holds at the individual publisher level, with some variation in grade and among publishers and one exception (The Washington Post loses representation between OpenWebText and OpenWebText2):

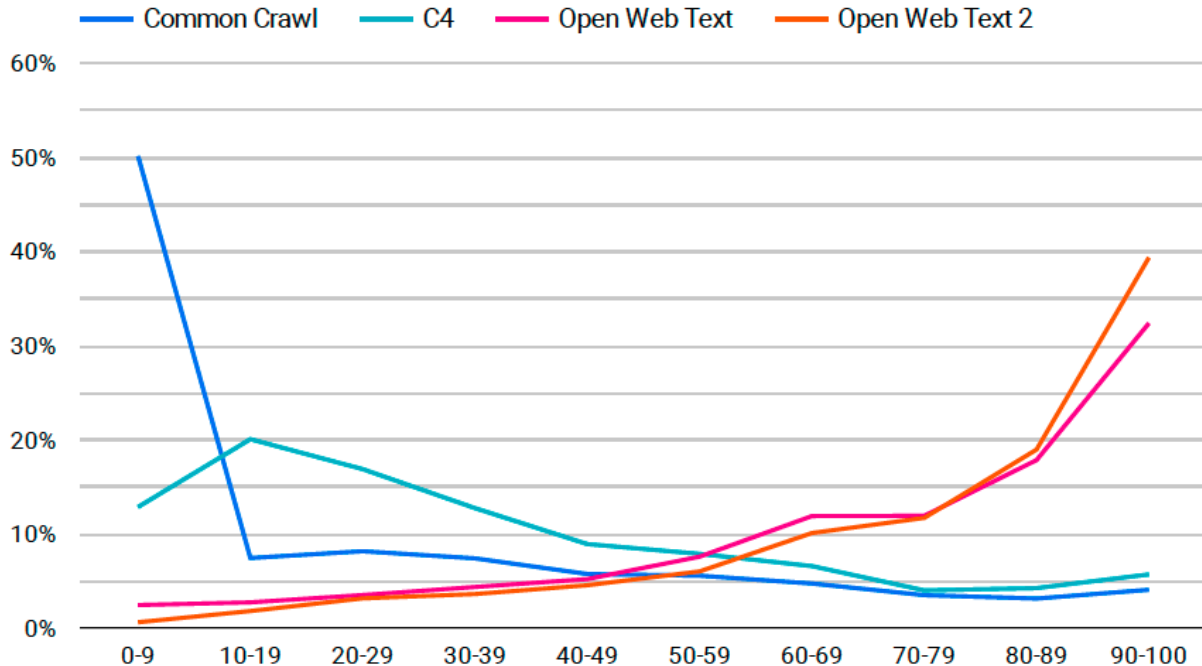


Publisher	Common Crawl	C4	OpenWebText	WebText top 1000	OpenWebText 2
News Corp	0.040%	0.132%	1.067%	1.131%	1.929%
The NY Times Co.	0.005%	0.048%	1.242%	1.737%	1.287%
Advance	0.030%	0.108%	0.940%	1.222%	1.236%
Vox Media	0.007%	0.023%	0.743%	0.905%	1.210%
Gannett	0.121%	0.346%	0.936%	0.910%	1.142%
Axel Springer	0.053%	0.053%	0.562%	0.791%	0.809%
Ziff Davis	0.033%	0.134%	0.539%	0.877%	0.656%
Alden Capital	0.018%	0.123%	0.617%	0.814%	0.647%
Hearst	0.048%	0.130%	0.504%	0.431%	0.573%
Penske	0.016%	0.068%	0.343%	0.552%	0.561%
Washington Post	0.002%	0.022%	0.933%	1.645%	0.556%
Buzzfeed, Inc.	0.007%	0.042%	0.825%	1.405%	0.551%
Future	0.031%	0.154%	0.296%	0.383%	0.420%
IAC	0.022%	0.134%	0.311%	0.602%	0.387%
Bustle Digital Group	0.006%	0.039%	0.056%	0.064%	0.082%
Total	0.441%	1.555%	9.913%	13.470%	12.047%

Our analysis of the top 1,000 domains in WebText disclosed by OpenAI shows that the publishers have similar shares of the URLs in that partial set. We include it in the table above for that reason, but exclude it from the graph because the more limited scope of that dataset obfuscates the trend shown by the other fully public datasets.

C. Percentage Contribution of URLs by Range of Domain Authority

As the LLM companies intended, dataset curation dramatically increases the quality (measured by Domain Authority) of a corpus. We compiled DA data from Moz for all URLs in Common Crawl, OpenWebText, OpenWebText2, and C4,⁴⁹ and graphed the distributions, discretized into ranges of 10 points of Domain Authority:



Visualized in this way, the URL DA distribution profiles of Common Crawl, C4 and the OpenWebText sets are informative. Common Crawl and C4 (impartially cleaned Common Crawl), on the one hand, and OpenWebText and OpenWebText2, on the other, fall into two different data shape categories.

Common Crawl and C4 are both left-skewed. Common Crawl peaks sharply in the 0–9 DA range, containing 50.2% of its domains, and drops off precipitously from there. C4 is more evenly distributed but remains moderately lumped to the left, gently peaking in the 10-19 DA range, at 20.1%.

The pattern is reversed for the curated corpora, in which high-DA URLs predominate. The discretized distributions of OpenWebText and OpenWebText2 DA scores are both right-skewed and similar in shape. Each climb steadily and then sharply as Domain Authority increases, peaking in the 90–100 DA range, at 32.4% and 39.4% for that range, respectively.⁵⁰

In sum, the uncuration datasets are weighted to low-quality, low-DA websites, while the curated datasets are heavily weighted to high-quality, high-DA websites. Our analysis indicates that the

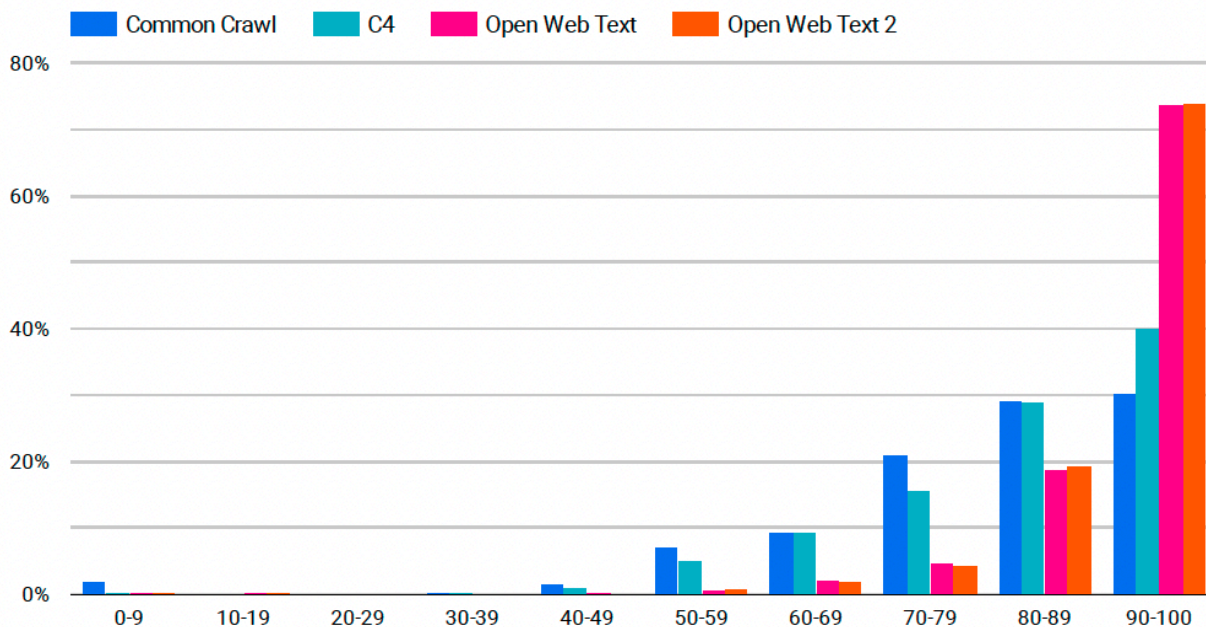
⁴⁹ We excluded the top 1,000 domains in WebText disclosed by OpenAI because that partial list of domains is inherently skewed towards more prolific (and potentially more authoritative) sources.

⁵⁰ Clustering of URLs at the low end absent curation was even more pronounced in non-discretized distributions both before and after including all domains' DA scores in Moz's 1.1 billion domain database. Both Common Crawl and Moz's all-web data are heavily skewed to the left.

process of curation of LLM training corpora favors high Domain Authority websites, the metric best representing quality of web content.

D. Connecting Correlations

The commercial portfolio publishers largely produce high-DA content. When we limited the set of domains to those publishers' domains, we saw that they provide little to no content from domains with Domain Authority scores of 50 or lower, even within the uncured corpora. Their domains cluster primarily into the top 90–100 DA range.



This observation is consistent with the correlations that we have observed across the LLM datasets, that increasing curation is correlated with (1) more commercial publisher sourcing and (2) higher Domain Authority. The former drives the latter because the commercial publishers we studied publish high-Domain Authority websites. Each of these trends reflects decisions made by LLM companies to prioritize high-quality web text datasets in training LLMs, resulting in revolutionary technological breakthroughs driving enormous value for those companies.

VI. Litigation

Numerous content owners (including publishers) have sued OpenAI, Microsoft, Alphabet/Google, Anthropic, and other generative AI companies, alleging copyright infringement and related claims. The publisher plaintiffs have supported their claims with the public evidence that their content—and high-quality content generally—was significantly present and disproportionately weighted in LLM training corpora including CC, C4, OpenWebText, and OpenWebText2.

The New York Times Company, in its copyright infringement lawsuit against OpenAI and Microsoft, pointed to its high rankings for documents (5th) and volume (15th) in WebText. It emphasized the materiality of WebText2 to GPT-3 (having 22% weight but only 4% of the tokens of the training data) and found that New York Times URLs accounted for 1.23% of OpenWebText2, which the New York Times treated as a proxy for WebText2, as we do. The

New York Times also referred to OpenAI's statement in the GPT-3 paper that higher-quality datasets were sampled most frequently during training.⁵¹

The New York Daily News, the Chicago Tribune, and eight other newspapers (all owned by Alden Global Capital), sued OpenAI and Microsoft for copyright infringement in April 2024, and this case has been consolidated with the New York Times case. Even alone, the Alden case is a significant AI litigation. The newspaper plaintiffs counted their documents in WebText, counted tokens in C4 as a proxy for Common Crawl, and referred to OpenAI's statement about the primacy of quality. They presumed that OpenAI and Microsoft have continued to use WebText and WebText2 to train models after GPT-3.⁵²

The Center for Investigative Reporting, in its June 27, 2024 Complaint against OpenAI and Microsoft, zeroed in on WebText's use of the Dragnet and Newspaper text extraction algorithms. These algorithms are known and designed to separate the body text of an article from other parts of a web page, including—meaningfully—terms of use links and copyright notices. Copyright notices and terms of use can constitute “copyright management information” or “CMI,” which is unlawful to remove from a copyrighted work. The Center for Investigative Reporting therefore made claims for removal of CMI and subsequent distribution in violation of the federal Digital Millennium Copyright Act against OpenAI and Microsoft, in addition to copyright infringement claims.⁵³

VII. Conclusion

LLM company training data disclosures—largely dating to earlier, pure-research periods of the technology's evolution—and analysis of public training datasets show long-running exploitation of high-quality publisher content (extremely lucrative for the LLM companies) and imply lost licensing revenue from some of the world's most highly-valued companies.

Publishers can quantify LLM company appropriation of their content and evidence its fundamental importance to those companies' breakthrough LLMs and, by inheritance, succeeding frontier models. The association between publishers' high quality web content and LLM value is manifest in the data. Additional relationships between content quality and LLM capabilities should be studied as researchers seek to understand how LLMs actually work and the nature of their emergent language capabilities.⁵⁴

⁵¹ Complaint, *The N.Y. Times co. vs. Microsoft Corp.*, https://nytc-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf, ¶¶ 85, 87, 90 (citing Brown, et al.).

⁵² Complaint, *Daily News, LP v. Microsoft Corp.*, <https://s3.documentcloud.org/documents/24628242/daily-news-lp-et-al-v-microsoft-corporation-et-a-34-32-85-complaint.pdf>, ¶¶ 82–89.

⁵³ Complaint, *The Center for Investigative Reporting, Inc. v. OpenAI, Inc.*, https://www.motherjones.com/wp-content/uploads/2024/06/CIR_Lawsuit_Against-OpenAI_06.27.24.pdf, ¶¶ 58–67, 135–159. The New York Daily News complaint also made this argument.

⁵⁴ For a recent breakthrough in this “interpretability” work see Templeton, et al., “Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet,” <https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html> (2024).

Authors

George Wukoson is a media and technology business lawyer with expertise in copyright and internet content. He is Ziff Davis' Lead Attorney on AI Matters and is General Counsel of Ziff Davis' Health and Wellness division. George oversees AI governance and related initiatives and advises cross-departmental leaders on AI. Prior to Ziff Davis, George was a litigator and counselor at Davis Wright Tremaine—specializing in copyright and trademark work in the digital media and technology spaces—and Dechert. George sits on the Copyright and Literary Property Committee of the New York City Bar Association and on its AI and Numerical and Business Information subcommittees. George holds a Bachelor of Arts degree from Dartmouth College and a Juris Doctor degree from New York University.

Joey Fortuna is Chief Technology Officer of Ziff Davis, where he oversees corporate IT initiatives, policies and practices, and facilitates cross-company technology initiatives. A central figure in driving innovation and developing proprietary technology across Ziff Davis' media businesses, Joey has more than 30 years of experience in technical leadership, software engineering methodology, systems architecture, data services, analytics, product design and support. Prior to Ziff Davis, Joey was Senior Vice President of Engineering at VINDICO Group, VP of Engineering at Quigo, and co-founder and Vice President of Technology at Pseudo Programs, Inc. Joey holds seven patents for machine learning algorithms and devices and is a member of the Advisory Council of Glasswing Ventures, a venture capital firm specializing in AI startups. Joey holds a Bachelor's degree from Georgetown University and a Master's of Science in Machine Learning from Columbia University.

Appendix 1

Domain names of publishers used for LLM training datasets queries:

- Advance (Condé Nast, Advance Local)
 - acbj.com
 - al.com
 - alabamaeducationlab.org
 - allure.com
 - architecturaldigest.com
 - arstechnica.com
 - bizequity.com
 - bizjournals.com
 - bonappetit.com
 - cleveland.com
 - cntraveler.com
 - cntraveller.com
 - condenast.com
 - epicurious.com
 - glamour.com
 - glamourmagazine.co.uk
 - gourmet.com
 - gq-magazine.co.uk
 - gq.com
 - gulflive.com
 - hemmings.com
 - hereisoregon.com
 - houseandgarden.co.uk
 - ironman.com
 - jerseysbest.com
 - johansens.com
 - lacucinaitaliana.com
 - leadersinsport.com
 - lehighvalleylive.com
 - lonestarlive.com
 - masslive.com
 - mlive.com
 - newyorker.com
 - newyorkupstate.com
 - newzoo.com
 - nj.com
 - nyup.com
 - oregonlive.com
 - pennlive.com
 - pennstudios.media
 - peopleofalabama.com
 - pitchfork.com
 - reckon.news
 - self.com
 - silive.com
 - sportsbusinessjournal.com
 - syracuse.com
 - tatler.com
 - teenvogue.com
 - them.us
 - thisisalabama.org
 - turnitin.com
 - vanityfair.com
 - vogue.co.uk
 - vogue.com
 - voguebusiness.com
 - wearepop.com
 - wired.com
 - wired.com
 - worldofinteriors.com
- Alden Global Capital (Tribune Publishing, MediaNews Group)
 - advocate-news.com
 - altdaily.com
 - bocopreps.com
 - bostonherald.com
 - broomfieldenterprise.com
 - buffzone.com
 - canoncitydailyrecord.com
 - chicagomag.com
 - chicagotribune.com
 - chicoer.com
 - citizensvoice.com
 - coloradohometownweekly.com
 - m
 - courant.com
 - dailybreeze.com
 - dailybulletin.com
 - dailycamera.com
 - dailydemocrat.com
 - dailyfreeman.com
 - dailylocal.com
 - dailynews.com

- dailypress.com
- dailytribune.com
- delcotimes.com
- denverpost.com
- eastbaytimes.com
- fortmorganimes.com
- greeleytribune.com
- journal-advocate.com
- lamarledger.com
- lowellsun.com
- macombdaily.com
- marinij.com
- mcall.com
- medianewsgroup.com
- mercurynews.com
- metromix.com
- militarynews.com
- montereyherald.com
- morningjournal.com
- mytowncolorado.com
- napervillemagazine.com
- nashobavalleyvoice.com
- news-herald.com
- nydailynews.com
- ocregister.com
- orlandosentinel.com
- orovillemr.com
- paradisepost.com
- pasadenastarnews.com
- pilotonline.com
- pottsmmerc.com
- pressenterprise.com
- presstelegram.com
- raisedintherockies.com
- readingeagle.com
- record-bee.com
- redbluffdailynews.com
- redlandsdailyfacts.com
- remindernews.com
- reporterherald.com
- republicanherald.com
- sandiegouniontribune.com
- santacruzsentinel.com
- saratogian.com
- sbsun.com
- sentinelandenterprise.com
- sgvtribune.com
- standardspeaker.com
- sun-sentinel.com
- themorningsun.com
- theoaklandpress.com
- thereporter.com
- thereporteronline.com
- thetimes-tribune.com
- times-standard.com
- timescall.com
- timesherald.com
- timesheraldonline.com
- trentonian.com
- tribpub.com
- tribunecontentagency.com
- troyrecord.com
- twincities.com
- ukiahdailyjournal.com
- virginiamedia.com
- whittierdailynews.com
- willitsnews.com
- Axel Springer
 - antenne.de
 - aufeminin.com
 - auto-bild.ro
 - auto-swiat.pl
 - auto.cz
 - autobazar.eu
 - autobild.de
 - autobild.es
 - autoexpress.co.uk
 - autohaus24.de
 - autojournal.fr
 - autoplus.fr
 - autoweek.nl
 - awin.com
 - axel-springer-akademie.de
 - axelspringer-syndication.de
 - axelspringer.com
 - axelspringer.de
 - azet.sk

- bams.de
- belvilla.com
- beobachter.ch
- bild.de
- blau-magazin.de
- blic.rs
- blikk.hu
- bonial.com
- bonial.fr
- businessinsider.com
- businessinsider.com.au
- businessinsider.com.pl
- businessinsider.de
- businessinsider.es
- businessinsider.in
- businessinsider.jp
- businessinsider.nl
- bz-berlin.de
- bz.de
- candidatemanager.net
- careerjunction.co.za
- carwale.com
- casamundo.com
- computerbild.de
- computerhoy.com
- ein-herz-fuer-kinder.de
- emarketer.com
- europeanvoice.com
- fakt.pl
- ffh.de
- ffn.de
- finanz.ru
- finanzen.ch
- finanzen.net
- fitbook.de
- forbes.pl
- forbes.ru
- geo.ru
- gofeminin.de
- good.co
- gruenderszene.de
- handelszeitung.ch
- hobbyconsolas.com
- homeday.de
- housell.com
- hy.co
- idealo.co.uk
- idealo.de
- idealo.fr
- ikiosk.de
- immonet.de
- immowelt.de
- insider-inc.com
- insider.com
- insiderintelligence.com
- irishjobs.ie
- jakdojade.pl
- jobs.ie
- jobsite.co.uk
- kaufda.de
- komputerswiat.pl
- ladenzeile.de
- logic-immo.com
- lonny.com
- mediaimpact.de
- medonet.pl
- meilleursagents.com
- metal-hammer.de
- musikexpress.de
- n24.de
- newsweek.pl
- nin.co.rs
- noizz.pl
- ofeminin.pl
- onet.pl
- opineo.pl
- ozy.com
- pclab.pl
- periodismodelmotor.com
- playpc.pl
- plejada.pl
- pme.ch
- pnet.co.za
- politico.com
- politico.eu
- profession.hu
- protocol.com
- przegladsportowy.pl

- radiohamburg.de
- rollingstone.de
- saongroup.com
- schweizerbank.ch
- seloger.com
- shareasale.com
- shopalike.dk
- shopalike.es
- shopalike.fi
- shopalike.fr
- shopalike.nl
- shopalike.se
- shopalike.sk
- skapiec.pl
- sportauto.fr
- stepstone.at
- stepstone.be
- stepstone.com
- stepstone.de
- stepstone.fr
- stepstone.nl
- students.ch
- stylebistro.com
- techbook.de
- topgear.es
- totaljobs.com
- transfermarkt.de
- travelbook.de
- upday.com
- usgang.ch
- vod.pl
- wams.de
- welt.de
- zimbio.com
- zumi.pl
- zumper.com
- Bustle Digital Group
 - bdg.com
 - bustle.com
 - elitedaily.com
 - fatherly.com
 - inverse.com
 - mic.com
- nylon.com
- romper.com
- scarymommy.com
- thezoereport.com
- wmagazine.com
- BuzzFeed, Inc.
 - buzzfeed.com
 - buzzfeednews.com
 - firstwefeast.com
 - huffingtonpost.co.uk
 - huffingtonpost.com
 - huffpost.com
 - tasty.co
- Comcast NBCUniversal
 - bravotv.com
 - cnbc.com
 - cnbc.comworld
 - dreamworks.com
 - eonline.com
 - fandango.com
 - focusfeatures.com
 - golfnow.com
 - hayu.com
 - illumination.com
 - msnbc.com
 - nbc.com
 - nbcbayarea.com
 - nbcboston.com
 - nbcchicago.com
 - nbccconnecticut.com
 - nbcdfw.com
 - nbclosangeles.com
 - nbcmiami.com
 - nbcnews.com
 - nbcnewyork.com
 - nbcolympics.com
 - nbcphiladelphia.com
 - nbcsandiego.com
 - nbcsports.com
 - nbcsportsbayarea.com
 - nbcsportsnext.com
 - nbcsportsphiladelphia.com

- nbcuproductions.com
- nbcwashington.com
- oxygen.com
- peacocktv.com
- rottentomatoes.com
- rwsentosa.com
- sny.tv
- sportsengine.com
- syfy.com
- telemundo.com
- telemundo20.com
- telemundo31.com
- telemundo33.com
- telemundo40.com
- telemundo47.com
- telemundo48elpaso.com
- telemundo49.com
- telemundo51.com
- telemundo52.com
- telemundo62.com
- telemundoareadelabahia.com
- telemundoarizona.com
- telemundochicago.com
- telemundodallas.com
- telemundodenver.com
- telemundofresno.com
- telemundohouston.com
- telemundolasvegas.com
- telemundonuevainglaterra.com
- telemundopr.com
- telemundosanantonio.com
- telemundoutah.com
- telemundowashingtondc.com
- universalbeijingresort.com
- universalkids.com
- universalorlando.com
- universalpictures.co.uk
- universalpictures.com
- universalproductsexperiences.com
- universalstudioshollywood.com

- uphe.com
- usanetwork.com
- usj.co.jp

- Future plc

- 5gradar.com
- actualtechmedia.com
- advnture.com
- anandtech.com
- androidcentral.com
- apcmag.com
- avnetwork.com
- avnetworknation.com
- bchalloffame.com
- bikeperfect.com
- brandimpactawards.com
- broadcastingcable.com
- channelpro.co.uk
- cinemablend.com
- cloudpro.co.uk
- coachmag.co.uk
- countrylife.co.uk
- creativebloq.com
- creativeplanetnetwork.com
- cyclingnews.com
- cyclingweekly.com
- decanter.com
- digitalcameraworld.com
- falltvevents.com
- fitandwell.com
- fourfourtwo.com
- future.swoogo.com
- futureevents.uk
- gamesradar.com
- gardeningetc.com
- gardeningknowhow.com
- getcomputeractive.co.uk
- getprice.com.au
- gocompare.com
- goldenjoystickawards.com
- golf-monthly.co.uk
- golfmonthly.com
- goodto.com
- guitarplayer.com

- guitarworld.com
- homebuilding.co.uk
- homebuildingshow.co.uk
- homesandgardens.com
- horseandhound.co.uk
- howitworksdaily.com
- idealhome.co.uk
- imore.com
- installation-international.com
- itpro.co.uk
- itproportal.com
- kiplinger.com
- laptopmag.com
- livescience.com
- livingetc.com
- livinggetcevents.com
- loudersound.com
- magazine.co.uk
- magazinesdirect.com
- marieclaire.co.uk
- mbr.co.uk
- mby.com
- mcnwonderwomen.com
- mixonline.com
- mixsoundforfilm.com
- mobilechoicework.com
- mobileindustryawards.com
- moderndad.com
- moneyweek.com
- mozo.com.au
- mrmobile.tech
- multichannel.com
- musicradar.com
- musicweek.com
- musicweektechsummit.com
- mw-womeninmusic.com
- myfavouritemagazines.co.uk
- myimperfectlife.com
- myvouchercode.co.uk
- nexttv.com
- pbo.co.uk
- pcgamer.com
- pcgamingshow.com
- petsradar.com
- photographyshow.com
- plotfinder.net
- practicalcaravan.com
- practicalmotorhome.com
- pricepanda.com
- prosoundnetwork.com
- radioworld.com
- realhomes.com
- residentialsystems.com
- rugbyworld.com
- sciencenature.theweekjunior.co.uk
- shortlist.com
- smartbrief.com
- space.com
- subscribe.pcpro.co.uk
- subscribe.techmags.com.au
- svconline.com
- t3.com
- techlearning.com
- technobuffalo.com
- technologyleadershipsummit.com
- techradar.com
- techradarchoiceawards.com
- thefield.co.uk
- theweek.co.uk
- theweekjunior.co.uk
- thrifter.com
- tomsguide.com
- tomshardware.com
- toptenreviews.com
- total911.com
- tvbeurope.com
- tvtechnology.com
- twice.com
- vertexconf.com
- wallpaper.com
- whatculture.com
- whathifi.com
- whattowatch.com
- whowhatwear.com
- windowscentral.com
- womanandhome.com

- womanmagazine.co.uk
- womansweekly.co.uk
- yachtingmonthly.com
- yachtingworld.com
- ybw.com
- Gannett
 - aberdeennews.com
 - allooadvertiser.com
 - amarillo.com
 - amestrib.com
 - andoveradvertiser.co.uk
 - app.com
 - ardrossanherald.com
 - argusleader.com
 - asianimage.co.uk
 - augustachronicle.com
 - autoexchange.co.uk
 - ayradvertiser.com
 - azcentral.com
 - banburycake.co.uk
 - barkinganddagenhampost.co.uk
 - barrheadnews.com
 - barryanddistrictnews.co.uk
 - basildonstandard.co.uk
 - basingstokegazette.co.uk
 - battlecreekenquirer.com
 - beaconjournal.com
 - becclesandbungayjournal.co.uk
 - bicesteradvertiser.net
 - blueridgenow.com
 - bordercountiesadvertizer.co.uk
 - bordertelegraph.com
 - borehamwoodtimes.co.uk
 - bournemouthecho.co.uk
 - bracknellnews.co.uk
 - braintreeandwithamtimes.co.uk
 - brentwoodlive.co.uk
 - bridgwatermercury.co.uk
 - bridportnews.co.uk
 - bromsgroveadvertiser.co.uk
 - buckscountycouriertimes.com
 - bucksfreepress.co.uk
 - bucyrustelegraphforum.com
 - burlingtoncountytimes.com
 - burlingtonfreepress.com
 - burnhamandhighbridgeweeklynnews.co.uk
 - burytimes.co.uk
 - businessignortheast.co.uk
 - caller.com
 - cambstimes.co.uk
 - campaignseries.co.uk
 - cantondailyledger.com
 - cantonrep.com
 - capecodtimes.com
 - carrickherald.com
 - centralfifetimes.com
 - chardandilminsternews.co.uk
 - cheboygannews.com
 - chelmsfordweeklynews.co.uk
 - chesterstandard.co.uk
 - chieftain.com
 - chillicothegazette.com
 - chorleycitizen.co.uk
 - cincinatti.com
 - cincinnati.com
 - citizen-times.com
 - cjonline.com
 - clactonandfringtongazette.co.uk
 - clarionledger.com
 - clydebankpost.co.uk
 - coloradoan.com
 - columbiadailyherald.com
 - columbiatribune.com
 - commercialappeal.com
 - coshoctontribune.com
 - cotswoldjournal.co.uk
 - countypress.co.uk
 - countytimes.co.uk
 - courier-journal.com
 - courierpostonline.com

- courierpress.com
- cravenherald.co.uk
- cumnockchronicle.com
- daily-jeff.com
- dailyamerican.com
- dailycomet.com
- dailycommercial.com
- dailyecho.co.uk
- dailyrecord.com
- dailyworld.com
- darlingtonandstocktontimes.co.uk
- delawareonline.com
- delmarvanow.com
- democratandchronicle.com
- denbighshirefreepress.co.uk
- derehamtimes.co.uk
- desertsun.com
- desmoinesregister.com
- dispatch.com
- dissmercury.co.uk
- dnj.com
- dorsetbeaches.co.uk
- dorsetecho.co.uk
- droitwichadvertiser.co.uk
- dudleynews.co.uk
- dumbartonreporter.co.uk
- dunfermlinepress.com
- dunmowbroadcast.co.uk
- eadt.co.uk
- ealingtimes.co.uk
- eastkilbrideconnect.co.uk
- eastlondonadvertiser.co.uk
- eastlothiancourier.com
- echo-news.co.uk
- edp24.co.uk
- ellwoodcityledger.com
- elpasotimes.com
- elystandard.co.uk
- enfieldindependent.co.uk
- enterpriseneeds.com
- eppingforestguardian.co.uk
- essex-media.co.uk
- essexbusinessawards.co.uk
- eveningnews24.co.uk
- eveningsun.com
- eveningtribune.com
- eveshamjournal.co.uk
- examiner-enterprise.com
- exmouthjournal.co.uk
- fakenhamtimes.co.uk
- falmouthpacket.co.uk
- fayobserver.com
- fdlreporter.com
- floridatoday.com
- forestryjournal.co.uk
- fosters.com
- freep.com
- freepressseries.co.uk
- gadsdentimes.com
- gainesville.com
- galesburg.com
- gastongazette.com
- gazette-news.co.uk
- gazetteandherald.co.uk
- gazetteherald.co.uk
- gazetteseries.co.uk
- glasgowtimes.co.uk
- goerie.com
- gosanangelo.com
- goupstate.com
- greatfallstribune.com
- greatyarmouthmercury.co.uk
- greenbaypressgazette.com
- greenocktelegraph.co.uk
- greenvilleonline.com
- guardian-series.co.uk
- hackneygazette.co.uk
- halesowennews.co.uk
- halsteadgazette.co.uk
- hamhigh.co.uk
- hampshirechronicle.co.uk
- harrowtimes.co.uk
- harwichandmanningtreestandard.co.uk
- hattiesburgamerican.com
- hawkcentral.com
- helensburghadvertiser.co.uk

- heraldmailmedia.com
- heraldnews.com
- heraldscotland.com
- heraldseries.co.uk
- heraldtimesonline.com
- heraldtribune.com
- herefordtimes.com
- hertsad.co.uk
- hexham-courant.co.uk
- hillingdontimes.co.uk
- hillsdale.net
- hollandsentinel.com
- houmatoday.com
- htrnews.com
- huntspost.co.uk
- hutchnews.com
- ilfordrecorder.co.uk
- ilkleygazette.co.uk
- impartialreporter.com
- in-cumbria.com
- indeonline.com
- independentmail.com
- indystar.com
- ipswichstar.co.uk
- irvinetimes.com
- islingtongazette.co.uk
- ithacajournal.com
- jacksonsun.com
- jacksonville.com
- jconline.com
- journalstandard.com
- jsonline.com
- keighleynews.co.uk
- kidderminstershuttle.co.uk
- kilburntimes.co.uk
- kitsapsun.com
- knoxnews.com
- knutsfordguardian.co.uk
- lancashiretelegraph.co.uk
- lancastereaglegazette.com
- lansingstatejournal.com
- largsandmillportnews.com
- lcsun-news.com
- ldnews.com
- leaderlive.co.uk
- ledburyreporter.co.uk
- leighjournal.co.uk
- lenconnect.com
- lincolncourier.com
- living-magazines.co.uk
- livingstondaily.com
- localberkshire.co.uk
- lohud.com
- lowestoftjournal.co.uk
- lubbockonline.com
- ludlowadvertiser.co.uk
- makeitgrateful.com
- maldonandburnhamstandard.co.uk
- malvern gazette.co.uk
- mansfieldnewsjournal.com
- marionstar.com
- marshfieldnews herald.com
- mcdonoughvoice.com
- messengernewspapers.co.uk
- metrowestdailynews.com
- midweekherald.co.uk
- milforddailynews.com
- milfordmercury.co.uk
- monroenews.com
- montgomeryadvertiser.com
- mpnnow.com
- mycentraljersey.com
- naplesnews.com
- newarkadvocate.com
- newforestpost.co.uk
- newhamrecorder.co.uk
- newportri.com
- news-journalonline.com
- news-leader.com
- news-press.com
- newsandstar.co.uk
- newschief.com
- news herald.com
- newsleader.com
- newsshopper.co.uk
- njherald.com
- northernfarmer.co.uk

- northjersey.com
- northnorfolknews.co.uk
- northsomerstimes.co.uk
- northwaleschronicle.co.uk
- northwalespioneer.co.uk
- northwichguardian.co.uk
- norwichbulletin.com
- nwemail.co.uk
- nwfdailynews.com
- oakridger.com
- ocala.com
- oklahoman.com
- onlineathens.com
- oxfordmail.co.uk
- pal-item.com
- palmbeachdailynews.com
- palmbeachpost.com
- patriotledger.com
- peeblesshirenews.com
- pekintimes.com
- penarthtimes.co.uk
- peterboroughmatters.co.uk
- petoskeynews.com
- pinkun.com
- pjstar.com
- pnj.com
- poconorecord.com
- pontiacdailyleader.com
- portclintonnewsherald.com
- postcrescent.com
- poughkeepsiejournal.com
- press-citizen.com
- pressconnects.com
- prestwichandwhitefieldguide.co.uk
- prime-magazine.co.uk
- progress-index.com
- providencejournal.com
- publicopiniononline.com
- readingchronicle.co.uk
- record-courier.com
- Recordnet.com
- RecordOnline.com
- redding.com
- redditchadvertiser.co.uk
- redhillandreigatelife.co.uk
- registerguard.com
- reporter-times.com
- reporternews.com
- rgj.com
- rhyljournal.co.uk
- richmondandtwickenhamtimes.co.uk
- romfordrecorder.co.uk
- romseyadvertiser.co.uk
- royston-crow.co.uk
- rrstar.com
- runcornandwidnesworld.co.uk
- saffronwaldenreporter.co.uk
- salina.com
- salisburyjournal.co.uk
- savannahnow.com
- sctimes.com
- seacoastonline.com
- sheboyganpress.com
- shelbystar.com
- shreveporttimes.com
- sidmouthherald.co.uk
- sj-r.com
- sloughobserver.co.uk
- somersetcountygazette.co.uk
- sooeveningnews.com
- southbendtribune.com
- southcoasttoday.com
- southendstandard.co.uk
- southernkitchen.com
- southwalesargus.co.uk
- southwalesguardian.co.uk
- southwestfarmer.co.uk
- stalbansreview.co.uk
- starcourier.com
- stargazette.com
- starnewsonline.com
- statesman.com
- statesmanjournal.com
- staugustine.com
- stevenspointjournal.com

- sthensstar.co.uk
- stourbridgenews.co.uk
- stroudnewsandjournal.co.uk
- sturgisjournal.com
- surreycomet.co.uk
- swanageandwarehamvoice.co.uk
- swindonadvertiser.co.uk
- swtimes.com
- tallahassee.com
- tauntongazette.com
- tcpalm.com
- telegram.com
- tennessean.com
- tewkesburyadmag.co.uk
- the-daily-record.com
- the-leader.com
- the-review.com
- theadvertiser.com
- theargus.co.uk
- theboltonnews.co.uk
- thecalifornian.com
- thecomet.net
- thedailyjournal.com
- thedailyreporter.com
- thegardnernews.com
- thegleaner.com
- theintell.com
- thelancasterandmorecambeitizen.co.uk
- theleafchronicle.com
- theledger.com
- thenational.scot
- thenews-messenger.com
- thenewsstar.com
- thenorthernecho.co.uk
- thenorthwestern.com
- theoldhamtimes.co.uk
- thepublicopinion.com
- therecordherald.com
- thescottishfarmer.co.uk
- thespectrum.com
- thestarpress.com
- thetelegraphandargus.co.uk
- thetfordandbrandontimes.co.uk
- thetimesherald.com
- thetottenhamindependent.co.uk
- thetowntalk.com
- thewestmorlandgazette.co.uk
- thewestonmercury.co.uk
- thisislancashire.co.uk
- thisislocallondon.co.uk
- thisisoxfordshire.co.uk
- thisisthewestcountry.co.uk
- thisiswiltshire.co.uk
- thurrockgazette.co.uk
- times-gazette.com
- times-series.co.uk
- timesandstar.co.uk
- timesonline.com
- timesrecordnews.com
- timesreporter.com
- timestelegram.com
- tivysideadvertiser.co.uk
- tmnews.com
- tricountyindependent.com
- troontimes.com
- tucson.com
- tuscaloosaneews.com
- usatoday.com
- uticaod.com
- vcstar.com
- visaliatimesdelta.com
- vvdailypress.com
- walesfarmer.co.uk
- wandsworthguardian.co.uk
- warringtonguardian.co.uk
- watfordobserver.co.uk
- wattonandswaffhamtimes.co.uk
- wausaudailyherald.com
- wearevoice.co.uk
- westerntelegraph.co.uk
- wharfedaleobserver.co.uk
- whitchurchherald.co.uk
- whitehavennews.co.uk

- whtimes.co.uk
 - wiltsglosstandard.co.uk
 - wiltshirebusinessonline.co.uk
 - wiltshiretimes.co.uk
 - wimbledonguardian.co.uk
 - windsorobserver.co.uk
 - winsfordguardian.co.uk
 - wirralglobe.co.uk
 - wisbechstandard.co.uk
 - wisconsinrapidstribune.com
 - witneygazette.co.uk
 - worcesternews.co.uk
 - wymondhamandattleboroughmercury.co.uk
 - ydr.com
 - yeovilexpress.co.uk
 - yorkpress.co.uk
 - yourlocalguardian.co.uk
 - zanesvilletimesrecorder.com
- Hearst
 - 4029tv.com
 - 98online.com
 - abc-7.com
 - arcwcrew.com
 - autoweek.com
 - beaumontenterprise.com
 - bestproducts.com
 - bicycling.com
 - bigrapidsnews.com
 - biography.com
 - blaze.tv
 - caranddriver.com
 - chron.com
 - cosmopolitan.com
 - countryliving.com
 - crimeandinvestigationnetwork.com
 - ctpost.com
 - delish.com
 - elle.com
 - elledecor.com
 - esquire.com
 - expressnews.com
 - fdbhealth.com
 - firstfinds.com
 - fitchratings.com
 - fitchsolutions.com
 - fyi.tv
 - goodhousekeeping.com
 - greenwichtime.com
 - harpersbazaar.com
 - hearstmags.com
 - housebeautiful.com
 - houstonchronicle.com
 - kcci.com
 - kcra.com
 - ketv.com
 - kingfeatures.com
 - kmbc.com
 - koat.com
 - koco.com
 - ksbw.com
 - lmtonline.com
 - manisteenews.com
 - marieclaire.com
 - menshealth.com
 - michigansthum.com
 - middletownpress.com
 - mor-tv.com
 - mrt.com
 - myjournalcourier.com
 - mynbc5.com
 - myplainview.com
 - mysanantonio.com
 - nbc-2.com
 - newstimes.com
 - nhregister.com
 - oprahdaily.com
 - ourmidland.com
 - popularmechnics.com
 - prevention.com
 - redbookmag.com
 - registercitizen.com
 - roadandtrack.com
 - runnersworld.com
 - seattlepi.com
 - seventeen.com

- sfchronicle.com
- sfgate.com
- stamfordadvocate.com
- thehour.com
- theintelligencer.com
- thepioneerwoman.com
- thetelegraph.com
- timesunion.com
- townandcountrymag.com
- veranda.com
- wapt.com
- wbal.com
- wbaltv.com
- wcvb.com
- wdsu.com
- weekand.com
- wesh.com
- wgal.com
- wisn.com
- wjcl.com
- wlky.com
- wlwt.com
- wmtw.com
- wmur.com
- womansday.com
- womenshealthmag.com
- wpbf.com
- wptz.com
- wtae.com
- wvtm13.com
- wxii12.com
- wyff4.com
- A+E Networks (JV 50% owned Hearst; attributed to Hearst)
 - aenetworks.com
 - aetv.com
 - history.com
 - mylifetime.com
- IAC (Dotdash Meredith, Angi Inc., Search, and Emerging & Other divisions)
 - about.com
 - agriculture.com
 - allpeoplequilt.com
 - allrecipes.com
 - angi.com
 - ask.com
 - askmediagroup.com
 - bestlifeonline.com
 - bhg.com
 - bodynetwork.com
 - brides.com
 - byrdie.com
 - care.com
 - celebwell.com
 - cookinglight.com
 - craftjack.com
 - dailypaws.com
 - eatingwell.com
 - eatthis.com
 - ew.com
 - familycircle.com
 - fitnessmagazine.com
 - foodandwine.com
 - handy.com
 - health.com
 - hellogiggles.com
 - homeadvisor.com
 - homestars.com
 - instyle.com
 - investopedia.com
 - life.com
 - lifewire.com
 - liquor.com
 - magazines.com
 - marthastewart.com
 - midwestliving.com
 - mybuilder.com
 - mydomaine.com
 - myhammer.com
 - mywedding.com
 - parents.com
 - people.com
 - peopleenespanol.com
 - realage.com
 - realsimple.com

- seriouseats.com
- shape.com
- simplyrecipes.com
- southernliving.com
- techboomers.com
- thebalance.com
- thebalancecareers.com
- thebalancemoney.com
- thebalancesmb.com
- thedailybeast.com
- thespruce.com
- thesprucecrafts.com
- thespruceeats.com
- thesprucepets.com
- thoughtco.com
- travaux.com
- travelandleisure.com
- treehugger.com
- tripsavvy.com
- verywell.com
- verywellfamily.com
- verywellfit.com
- verywellhealth.com
- verywellmind.com
- vivian.com
- woodmagazine.com
- News Corp
 - adelaidenow.com.au
 - aetv.com.au
 - australiannewschannel.com.au
 - barrons.com
 - bestrecipes.com.au
 - binge.com.au
 - bodyandsoul.com.au
 - boxsetschannel.com.au
 - couriermail.com.au
 - dailytelegraph.com.au
 - decider.com
 - delicious.com.au
 - dowjones.com
 - eurekareport.com.au
 - factiva.com
 - fncstatic.com
 - fnlondon..com
 - fox8.com
 - foxbusiness.com
 - foxclassics.com.au
 - foxnews.com
 - foxnewsradio.com
 - foxshowcase.com.au
 - foxsports.com
 - foxsports.com.au
 - foxtel.com.au
 - foxtelmovies.com.au
 - ggq.com.au
 - harpercollins.com
 - heraldsun.com.au
 - hipages.com.au
 - historychannel.com.au
 - homelife.com.au
 - hubbl.com.au
 - investors.com
 - kayosports.com.au
 - Kidspot.com.au
 - lifestyle.com.au
 - lifestylefood.com.au
 - mainevent.com.au
 - mansionglobal.com
 - marketwatch.com
 - move.com
 - news.co.uk
 - news.com.au
 - newscorp.com
 - newscorpaustralia.com
 - northshoretimes.com.au
 - ntnews.com.au
 - nypost.com
 - odds.com.au
 - onebigswitch.com.au
 - pagesix.com
 - penews.com
 - punters.com.au
 - racenet.com.au
 - realestate.com.au
 - realtor.com
 - skynews.com.au

- storyful.com
- talk.tv
- talksport.com
- the-tls.co.uk
- theaustralian.com.au
- themercury.com.au
- thesun.co.uk
- thesundaymail.com.au
- thesundaytimes.co.uk
- thetimes.co.uk
- times.radio
- tips.com.au
- virginradio.co.uk
- vogue.com.au
- wsj.com
- wsj.net
- Penske Media Corporation
 - acmcountry.com
 - ampav.com
 - artforum.com
 - artnews.com
 - atxtv.co
 - atxtv.com
 - bgr.com
 - billboard.com
 - deadline.com
 - dickclark.com
 - footwearnnews.com
 - goldderby.com
 - goldenglobes.com
 - hollywoodreporter.com
 - indiewire.com
 - la3c.com
 - lifeisbeautiful.com
 - luminatedata.com
 - robbreport.com
 - rollingstone.com
 - sheknows.com
 - shemedia.com
 - sourcingjournal.com
 - sportico.com
 - spy.com
 - streamys.org
- stylecaster.com
- sxsw.com
- theamas.com
- theflowspace.com
- tvline.com
- variety.com
- vibe.com
- wwd.com
- Red Ventures
 - allconnect.com
 - bankrate.com
 - bestcolleges.com
 - bezzycb.com
 - bezzycopd.com
 - bezzidepression.com
 - bezzymbd.com
 - bezzymigraine.com
 - bezzyms.com
 - bezzypsa.com
 - bezzypsoriasis.com
 - bezzyra.com
 - bezzyt2d.com
 - confused.com
 - creditcards.com
 - greatist.com
 - healthgrades.com
 - healthline.com
 - lonelyplanet.com
 - medicalnewstoday.com
 - money.co.uk
 - mymove.com
 - perks.optum.com
 - plateapr.com
 - psychcentral.com
 - redventures.com
 - store.optum.com
 - thepointsguy.com
 - uswitch.com
- The New York Times Company
 - newyorktimes.com
 - nyt.com
 - nyt.net

- nytco.com
- nyti.ms
- nytimes.com
- theathletic.com
- thewirecutter.com
- wirecutter.com
- The Walt Disney Company
 - 20thcenturystudios.com
 - 6abc.com
 - abc.com
 - abc11.com
 - abc13.com
 - abc30.com
 - abc7.com
 - abc7chicago.com
 - abc7news.com
 - abc7ny.com
 - abcnews.com
 - abcnews.go.com
 - abcotvs.com
 - abcspark.ca
 - babytv.com
 - clubpenguin.com
 - dgepress.com
 - disney.co.uk
 - disney.com
 - disney.it
 - disney.pt
 - disneyabc.tv
 - disneyenglish.com
 - disneylachaine.ca
 - disneynow.com
 - disneyplus.com
 - disneystar.com
 - disneystudios.com
 - disneystudioshelp.com
 - dmdcentral.com
 - espn.com
 - foxmovies.com
 - freeform.com
 - fxnetworks.com
 - go.com
 - hollywoodrecords.com
 - hotstar.com
 - livewithkellyandmark.com
 - lucasfilm.com
 - marvel.com
 - moviesanywhere.com
 - natgeo.com
 - natgeomaps.com
 - nationalgeographic.com
 - nowtv.com.tr
 - patagonik.com.ar
 - radio.disney.com
 - radiodisney.com
 - searchlightpictures.com
 - starplus.com
 - tamronhallshow.com
 - telecine.com.br
 - tv.disney.es
- The Washington Post
 - arcxp.com
 - wapo.com
 - washingtonpost.com
 - washingtonpostlive.com
 - washingtonposttv.com
 - washpost.com
- Vox Media
 - acmepackingcompany.com
 - amazinavenue.com
 - arrowheadpride.com
 - aseaoofblue.com
 - athleticsnation.com
 - azdesertswarm.com
 - badlefthook.com
 - barcablaugranes.com
 - behindthesteelcurtain.com
 - beyondtheboxscore.com
 - bigblueview.com
 - bigcatcountry.com
 - blazersedge.com
 - bleedcubbieblue.com
 - bleedinggreennation.com
 - blessyouboys.com
 - bloggingtheboys.com

- bloodyyellow.com
- blueshirtbanter.com
- broadstreethockey.com
- bucsnation.com
- buffalorumblings.com
- burntorangenation.com
- cagesideseats.com
- celticsblog.com
- cincyjungle.com
- cornnation.com
- crawfishboxes.com
- curbed.com
- dailynorseman.com
- denverstiffs.com
- dukebasketballreport.com
- eater.com
- federalbaseball.com
- fieldgulls.com
- goldenstateofmind.com
- grubstreet.com
- halosheaven.com
- hammerandrails.com
- hogshaven.com
- letsgotribe.com
- lonestarmall.com
- lookoutlanding.com
- maizenbrew.com
- mavsmoneyball.com
- mccoveychronicles.com
- milehighreport.com
- mmafighting.com
- mmamania.com
- netsdaily.com
- ninersnation.com
- nowmedianetwork.co
- nymag.com
- onefootdown.com
- outsports.com
- overthemonster.com
- patspulpit.com
- pensburgh.com
- pinstripealley.com
- polygon.com
- popsugar.com
- postingandtoasting.com
- poundingtherock.com
- prideofdetroit.com
- racked.com
- rawcharge.com
- redreporter.com
- rollbamaroll.com
- royalsreview.com
- sbnation.com
- secondcityhockey.com
- silverandblackpride.com
- silverscreenandroll.com
- stanleycupofchowder.com
- talkingchop.com
- thecut.com
- thedodo.com
- thephinsider.com
- theringer.com
- theverge.com
- thrillist.com
- tomahawknation.com
- truebluela.com
- vivaelbirdos.com
- vox.com
- voxmedia.com
- vulture.com
- windycitygridiron.com
- Warner Bros. Discovery
 - adultswim.com
 - ahctv.com
 - animalplanet.com
 - asianfoodnetwork.com
 - bleacherreport.com
 - boingtv.it
 - boomerang.com
 - cartoonnetwork.com
 - cartoonnetwork.com
 - cinemax.com
 - cnn.com
 - cnnspanol.cnn.com
 - cookingchanneltv.com
 - dc.com
 - destinationamerica.com

- discovery.com
- discoverylife.com
- discoveryplus.com
- dmax.de
- eurosport.com
- foodnetwork.com
- giallotv.it
- golfdigest.com
- hbo.com
- hgtv.com
- hogarhgtv.com
- hotrod.com
- investigationdiscovery.com
- latamwbd.com
- max.com
- mondotv.jp
- motortrend.com
- nove.tv
- onefifty.com
- pogo.tv
- realtime.it
- roosterteeth.com
- sciencechannel.com
- tabichan.jp
- tbs.com
- tcm.com
- tlc.com
- tntdrama.com
- tntgo.tv
- travelchannel.com
- trutv.com
- tvn.pl
- warnerbros.com
- warnerbrodiscovery.dk
- warnerbrosgames.com
- warnertv.de
- wbd.com
- www.tooncast.tv
- babycenter.ca
- babycenter.com
- babycentre.co.uk
- blackfriday.com
- bphope.com
- campaigner.com
- castleconnolly.com
- classicgaming.com
- cnet.co
- cnet.co.kr
- cnet.co.uk
- cnet.com
- cnet.com.au
- cnet.de
- cnetfrance.fr
- couponcodes.com
- creditcardsexplained.com
- dailyom.com
- dealsofamerica.com
- diabetesdaily.com
- dicebreaker.com
- digitalfoundry.com
- digitalfoundry.net
- downdetector.ca
- downdetector.com
- download.com
- ekahau.com
- emmasdiary.co.uk
- encrypt.me
- ereceptionist.co.uk
- eurogamer.net
- everydayhealth.com
- evoice.com
- excelmicro.com
- extremetech.com
- familiesonline.co.uk
- fileplanet.com
- gamermetrics.com
- gamesindustry.biz
- gamespy.com
- gamestats.com
- gametab.com
- geek.com
- getinvisiblehand.com
- Ziff Davis
 - 1up.com
 - aberdeen.com
 - allestoringen.nl
 - askmen.com

- gizmos.com
- greenpixels.com
- healthcareers.com
- healthprofessionalacademy.co.uk
- hopetocope.com
- howlongtobeat.com
- humblebundle.com
- humblegames.com
- icontact.com
- ign.com
- ign.es
- ign.fr
- ignboards.com
- inspiredlearning.com
- ipvanish.com
- kickbox.com
- lifehacker.com
- line2.com
- livedrive.com
- loseit.com
- macfixit.com
- mapgenie.io
- mashable.com
- medpagetoday.com
- migraineagain.com
- moz.com
- myslumberryard.com
- news.com
- offers.com
- ookla.com
- pcmag.com
- pcmag.ru
- planetdiablo.com
- planetwolfenstein.com
- priceblink.com
- primece.com
- primeinc.org
- retailmenot.com
- rockpapershotgun.com
- rootmetrics.com
- safesendsoftware.com
- smartplanet.com
- smtp.com
- snowball.com
- speedtest.net
- spiceworks.com
- strongvpn.com
- sugarsync.com
- tdsgiftcards.com
- teamxbox.com
- techbargains.com
- theblackfriday.com
- ugo.com
- vg247.com
- vipre.com
- vouchercode.co.uk
- whattheyplay.com
- whattoexpect.com
- zd.net
- zdn.net
- zdn.co.uk
- zdn.com
- zdn.com.au
- zdn.de
- zdn.fr
- zdnasia.com
- zdperformancemarketing.com
- ziffdavis.com

Appendix 2

Comparison of domain evaluation techniques:

DOMAIN	MOZ	Ahrefs	ABS DIFFERENCE	ACTUAL DIFFERENCE
ign.com	92	88	4	4
mashable.com	93	91	2	2
cnet.com	93	92	1	1
healthline.com	91	92	1	-1
nytimes.com	95	94	1	1
thewirecutter.com	84	80	4	4
webmd.com	94	92	2	2
medscape.com	89	90	1	-1
chicagotribune.com	92	91	1	1
nydailynews.com	93	89	4	4
fandom.com	94	91	3	3
wikia.com	93	90	3	3
wsj.com	93	92	1	1
barrons.com	88	88	0	0
washingtonpost.com	94	93	1	1
countryliving.com	83	85	2	-2
esquire.com	90	88	2	2
ask.com	87	84	3	3
foodandwine.com	87	88	1	-1
glamour.com	89	86	3	3
gq.com	90	88	2	2
theverge.com	93	92	1	1
vulture.com	88	86	2	2
billboard.com	92	90	2	2
vibe.com	82	78	4	4

RANGE DIFFERENCE			2.04	1.64
capitolcorridor.org	49	58	9	-9
gunnerkrigg.com	50	52	2	-2
sixmorevodka.com	45	47	2	-2
epicwar.com	51	40	11	11
weatherwise.org	52	47	5	5
prototipo.io	47	58	11	-11
nav.com	56	75	19	-19
bancor.network	54	75	21	-21
advisornews.com	59	49	10	10
bitcoinexchangeguide.com	58	74	16	-16
RANGE DIFFERENCE			5.541666667	-1.375
wholesalesocksbulk.com	25	20	5	5
crip.to	21	12	9	9
neotv.cn	23	47	24	-24
implementingquantlib.con	20	13	7	7
yawningportal.org	25	11	14	14
arabic-for-nerds.com	25	29	4	-4
paniolotonewoods.com	23	12	11	11
passionspoon.com	21	6	15	15
dianebluemusic.com	21	12	9	9
touchstonelaw.com	25	9	16	16
atw.com	29	46	17	-17
warahma.com	3	0	3	3
mydragonflymoon.com	2	0	2	2
pavangupta.com	3	0	3	3
indiatour4u.com	6	0	6	6
enerludica.org	3	1	2	2
crystalinspections.com	5	4	1	1

monkeymoes.com	9	0	9	9
phantasybar.com	9	0	9	9
sitebin.com	7	0	7	7
heartlightconnection.com	7	0	7	7
RANGE DIFFERENCE			8.571428571	4.285714286
OVERALL DIFFERENCE			6.017857143	1.375