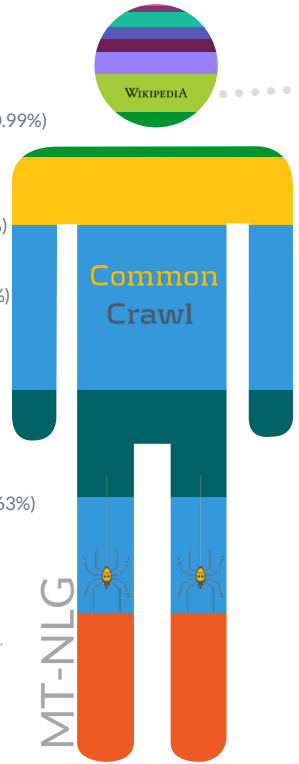
CONTENTS OF MEGATRON MT-NLG & THE PILE V1

- NIH ExPorter (papers) (0.2%)
- Gutenberg (books) (0.9%)
- CC-Stories (0.98%)
- BookCorpus2 (Libgen or similar) (0.99%)
- ArXiv (papers) (1.53%)
- GitHub (code) (1.79%)
- PubMed Abstracts (papers) (2.92%)
- Wikipedia (facts) (4.95%)
- Stack Exchange (discussion) (5.98%)
- RealNews (news) (8.88%)
- Pile-CC (www) (9.17%)
- CC-2020-50 (www) (12.66%)
- Books3 (Bibliotik tracker) (14.2%)
- CC-2021-04 (www) (15.22%)
- OpenWebText2 (Reddit links) (19.63%)



- Not to scale.
- Effective size by weighting (as % of total).
- Deduplication has been considered for Wikipedia.

Sources:
The Pile v1: https://arxiv.org/abs/2101.00027
C4: https://arxiv.org/abs/2104.08758
MT-NLG: see Microsoft & NVIDIA

Alan D. Thompson. October 2021.
https://lifearchitect.com.au/ai/

WebText (Reddit Submission Corpus)
HuffPost (news)
The New York Times (news)
BBC (news)
Twitter (discussion)
The Guardian (news)
The Washington Post (news)
and 4.3M+ more domains...

Common Crawl
(C4, cleaned/filtered, sorted by most tokens)
Google Patents (papers)
The New York Times (news)
Los Angeles Times (news)
The Guardian (news)
PLoS - Public Library of Science (papers)
Forbes (news)
HuffPost (news)
Patents.com – dead link (papers)
Scribd (books)
The Washington Post (news)
The Motley Fool (opinion)
InterPlanetary File System (mix)
Frontiers Media (papers)
Business Insider (news)
Chicago Tribune (news)
Booking.com (discussion)
The Atlantic (news)
Springer Link (papers)
Al Jazeera (news)
Kickstarter (discussion)
FindLaw Caselaw (papers)
National Center for Biotech Info (papers)
NPR (news)
and 90.9M+ more domains...

- Enron Emails (discussion) (0.14%)
- NIH ExPorter (papers) (0.3%)
- PhilPapers (papers) (0.38%)
- YoutubeSubtitles (movies) (0.6%)
- HackerNews (discussion) (0.62%)
- EuroParl (formal discussion) (0.73%)
- Books1/BookCorpus (Smashwords) (0.75%)
- Ubuntu IRC (discussion) (0.88%)
- DM Mathematics (papers) (1.24%)
- Wikipedia (facts) (1.53%)
- OpenSubtitles (movies) (1.55%)
- Gutenberg (books) (2.17%)
- PubMed Abstracts (papers) (3.07%)
- USPTO Background (papers) (3.65%)
- Stack Exchange (discussion) (5.13%)
- FreeLaw (papers) (6.12%)
- Github (code) (7.59%)
- ArXiv (papers) (8.96%)
- WebText (Reddit links) (10.01%)
- Books3 (Bibliotik tracker) (12.07%)
- PubMed Central (papers) (14.4%)
- Common Crawl (www) (18.11%)