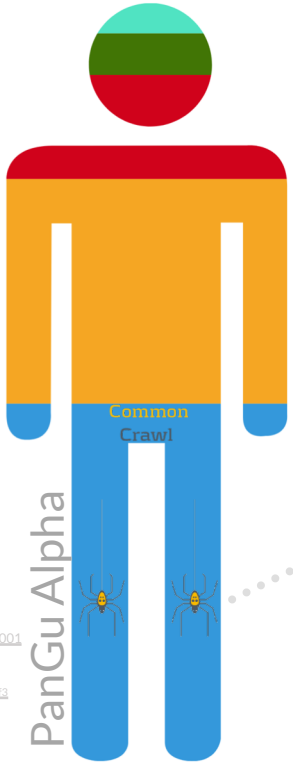
CONTENTS OF PANGU ALPHA & WUDAO 2.0

- e-Books (3.99%)
- Encyclopedia: (5.48%)
  - Baidu Baike (facts)
  - Sogou Baike (facts)
  - and more…
- News (13.72%)
- Public datasets: (29.65%)
  - DuReader (discussion)
  - BaiDuQA (discussion)
  - CAIL2018 (legal papers)
  - Sogou-CA (news)
  - and more…
- Common Crawl (www) (47.16%)



- Not to scale.
- Effective size by weighting (as % of total).
- WudaoCorpora 2.0 is 'best guess' only, no specifics available.

Sources:
PanGu Alpha: https://arxiv.org/abs/2104.12369
C4: https://arxiv.org/abs/2104.08758
Wudao1: https://doi.org/10.1016/j.aiopen.2021.06.001
Inverse prompting: https://arxiv.org/pdf/2103.10685.pdf
Wudao2 research report (BAAI conference notes): https://aminer.cn/research_report/60b628f030e4d5752f50def3
Alexa CN: https://alexa.com/topsites/countries/CN

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

Common Crawl
(C4, cleaned/filtered, sorted by most tokens)

Google Patents (papers)
Wikipedia English (facts)
Wikipedia Mobile English (facts)
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 more…

WuDaoCorpora 2.0
- Zhihu (discussion) (131GB)
- Baidu Baike (facts) (66.5GB)
- Sogou Baike (facts) (66.5GB)
- Baidu QA (discussion) (38GB)
- Other web pages*

*822M Web pages. Contents to be confirmed, 'best guess' only below, sorted by most visits.
- Tencent QQ (messenger)
- Sohu (news)
- Sina Weibo (discussion)
- Sina Corporation (news)
- Xinhua News Agency (news)
- Chinese Software Dev Network (discussion)
- Global Times (news)
- Tianya Club (discussion)
- 17ok.com (finance discussion)
- BabyTree (parenting discussion)
- CNBlogs (software discussion)
- 6Rooms (news)
- NetEase (discussion)
- Hunan Rednet (news)
- Bilibili (video discussion)
- Zhihu (discussion)
- and more...