EXHIBIT B

HIPI - We think that data is likely to be the main distinction between various AI labs over the next year, and where we think data could be found is actually quite a large competitive advantage. [redacted]

Some Context

Last year, we found that as we scale up models, we want to scale up parameters and data [redacted]

We've mostly kept up with compute needs on data so far, but two things have made this likely to become a major challenge going forward:

1. We think it's very likely that we will have [redacted] over the next couple years.
2. We have a new codenamed algorithmic approach that shifts us to require more data, by a factor of about [redacted] This is much more compute efficient, but if we use it (as seems likely for Teal), we will be pretty much maxed out on data.

There are 3 avenues to addressing this problem:

1. We can get more natural data from sources similar to what we have in the past (the focus of this doc)
2. We can generate synthetic data with our models. [redacted]
3. We can make algorithmic changes that let us do more repetitions of our data, or find other ways to spend more compute per token of data.

What type of data we need

We need human-written text or code, ideally with the following properties:

- Massive quantity. We currently have 10T tokens, and probably need to double this. The data team is 4 people. I think any source above [redacted] tokens ([redacted] of data) is worth looking into, but we really want to prioritize 1T+ scale data opportunities
- Long-context, meaning there are at least 300-100k words that are continuously causally related. E.g. a webpage or a book are good, a single isolated tweet is probably bad.
- Diverse topics - we can get this from multiple sources, but e.g. a trillion tokens about architecture wouldn't be good.
- [redacted]
- [redacted]

Our largest sources so far are [redacted] We think the [redacted] are higher quality. We've found in the past that high-quality data (e.g. books) makes a huge difference to our models, but we are now at the point where we really need quantity. For

Commented [1]: Do you have a rough estimate of how much data exists in the world? By major source? When will everyone just run out of natural data? [Author:Neerav Kingsland],[Creator:1297307460],[Date:8/12/2023 9:01:00 PM]

Commented [2]: Is synthetic data inevitably needed? [Author:Neerav Kingsland],[Creator:1297307460],[Date:8/12/2023 9:03:00 PM]

Commented [3]: We don't really know. My best guess [redacted] which is probably good enough for AGI. So synthetic data isn't inevitable, but it might be easier than other approaches.
1 total reaction
Neerav Kingsland reacted with 👍 at 2023-08-12 14:20 PM [Author:Nicholas Joseph],[Creator:1297307460],[Date:8/12/2023 9:12:00 PM]

example, we want to get another [redacted]50T tokens, and we believe that all the books are only [redacted] tokens.

Possible sources

Web Crawls

We've mostly looked into search engines here, but there are a few ways their goals differ from ours. In particular, search engines really care about having the latest version of commonly accessed webpages, whereas we care very little about freshness and much more about completeness

- [redacted]
- [redacted]
- [redacted]
- [redacted]
- [redacted]
- etc.
- CommonCrawl: This is what we currently use. [redacted], but anything they do is obviously given out to the whole community so it has substantial acceleration downsides
- [redacted]
- There is a list of other active crawlers here, we should maybe do outreach to each of them.
- I also think this is an area where someone with a less technical skillset might have an advantage.

Books

[redacted]

Other sources:

Commented [4]: Does global government produced data in aggregate hit a size threshold we'd want? Legislation, court rulings, regulations, reports, publicly filed documents (every 10K ever), clinical trials, all patents ever filed, etc [Author:Neerav Kingsland],[Creator:1297307460],[Date:8/12/2023 9:09:00 PM]

Commented [5]: Pretty sure the answer is yes. [Author:Nicholas Joseph],[Creator:1297307460],[Date:8/12/2023 9:13:00 PM]

Commented [6]: I imagine this is all publicly available. Should it be a tier one source opportunity? Could require a lot of manual labor. Maybe we could just FOIA the government for every document its ever produced :) [Author:Neerav Kingsland],[Creator:1297307460],[Date:8/12/2023 9:21:00 PM]

Commented [7]: Hm, my understanding is that for FOIA you need to be pretty specific about what you want? ... [1]

Commented [8]: We also have the "pile of law" ... [2]

Commented [9]: Sorry FOIA was a joke. ... [3]

Commented [10]: Do we have all publicly availab ... [4]

Commented [11]: https://guides.loc.gov/finding- ... [5]

Commented [12]: cc @kipply@anthropic.com on ... [6]

Commented [13]: https://www.regulations.gov/ ... [7]

Commented [14]: https://www.gpo.gov/ ... [8]

Commented [15]: https://www.parliament.uk/abo ... [9]

Commented [16]: https://www.ourcommons.ca/ ... [10]

Commented [17]: https://www.aph.gov.au/Parlia ... [11]

Commented [18]: https://libguides.bodleian.ox.a ... [12]

Commented [19]: [redacted] ... [13]

Commented [20]: [redacted]

Commented [21]: [redacted] ... [15]

[redacted]

- [redacted]
- [redacted] got us very excited at one point, but we couldn't crawl them effectively, and they didn't respond to our email requests.
- [redacted]
  - We have some of these, but I expect there are a ton on the internet that we haven't picked up.
  - [redacted]
- [redacted]
  - Probably too small, but I'm not sure. When I think of "places where people put effort into writing and that make it publicly available", blogs are one of the first things that comes to mind.
- [redacted]
  - [redacted]
  - [redacted]
  - [redacted]

Nick's best guess at a stack-ranked list of priorities

1. Talk to all the crawl companies
2. Talk to Twitter
3. See if there is anything that would convince [redacted] to give us books

Commented [29]: [redacted] [Author:Jared Kaplan],[Creator:1297307460],[Date:8/13/2023 7:11:00 AM]

Commented [30]: [redacted]

It seems like it might help if we get way more users next year though [Author:Nicholas Joseph],[Creator:1297307460],[Date:8/13/2023 3:07:00 PM]

Commented [31]: [redacted] Jared Kaplan],[Creator:1297307460],[Date:8/13/2023 5:01:00 PM]

Commented [32]: This sounds like a *very* Nova-shaped question [Author:Jared Kaplan],[Creator:1297307460],[Date:8/13/2023 7:12:00 AM]

Commented [33]: @amit@anthropic.com I maybe missed it, but didn't see these in your doc and they sound plausible / worth adding I think? [Author:Avital Balwit],[Creator:1297307460],[Date:12/8/2023 9:19:00 PM]

Commented [34]: Thanks for flagging, yes let me add so we have an exhaustive view of all our potential pathways [Author:Amit Desai],[Creator:1297307460],[Date:12/8/2023 9:49:00 PM]

Commented [35]: and hire for / do crawling [Author:Jared Kaplan],[Creator:1297307460],[Date:8/13/2023 7:15:00 AM]

Page 2: [1] Commented [7] Nicholas Joseph 8/12/2023 9:23:00 PM

Hm, my understanding is that for FOIA you need to be pretty specific about what you want?

What would be a first step here? I don't feel too clear on how to actually go about getting this. [Author:Nicholas Joseph],[Creator:1297307460],[Date:8/12/2023 9:23:00 PM]

Page 2: [2] Commented [8] Nicholas Joseph 8/12/2023 9:24:00 PM

We also have the "pile of law" dataset that has a bunch of court opinions and such, and I think we have all the patents too. So its more the internal government documents that we'd be looking for rather than ones that are already publicly available. [Author:Nicholas Joseph],[Creator:1297307460],[Date:8/12/2023 9:24:00 PM]

Page 2: [3] Commented [9] Neerav Kingsland 8/12/2023 9:30:00 PM

Sorry FOIA was a joke. [Author:Neerav Kingsland],[Creator:1297307460],[Date:8/12/2023 9:30:00 PM]

Page 2: [4] Commented [10] Neerav Kingsland 8/12/2023 9:30:00 PM

Do we have all publicly available stuff from sources like Edgar? https://www.sec.gov/edgar/searchedgar/companysearch Or here: https://www.govinfo.gov/bulkdata

I don't really know best first step.... [Author:Neerav Kingsland],[Creator:1297307460],[Date:8/12/2023 9:30:00 PM]

Page 2: [5] Commented [11] Neerav Kingsland 8/12/2023 9:35:00 PM

https://guides.loc.gov/finding-government-documents/portals-and-resources [Author:Neerav Kingsland],[Creator:1297307460],[Date:8/12/2023 9:35:00 PM]

Page 2: [6] Commented [12] Nicholas Joseph 8/13/2023 3:05:00 PM

cc @kipply@anthropic.com on these. I think we don't have them other than maybe some overlap in pile of law and they could be fairly large / easy to scrape [Author:Nicholas Joseph],[Creator:1297307460],[Date:8/13/2023 3:05:00 PM]

Page 2: [7] Commented [13] Neerav Kingsland 8/13/2023 8:22:00 PM

https://www.regulations.gov/ [Author:Neerav Kingsland],[Creator:1297307460],[Date:8/13/2023 8:22:00 PM]

Page 2: [8] Commented [14] Neerav Kingsland 8/13/2023 8:22:00 PM

https://www.gpo.gov/ [Author:Neerav Kingsland],[Creator:1297307460],[Date:8/13/2023 8:22:00 PM]

Page 2: [9] Commented [15] Neerav Kingsland 8/13/2023 8:26:00 PM

https://www.parliament.uk/about/how/publications/hansard/ [Author:Neerav Kingsland],[Creator:1297307460],[Date:8/13/2023 8:26:00 PM]

Page 2: [10] Commented [16] Neerav Kingsland 8/13/2023 8:27:00 PM

https://www.ourcommons.ca/documentviewer/en/35-2/house/hansard-index [Author:Neerav Kingsland],[Creator:1297307460],[Date:8/13/2023 8:27:00 PM]

Page 2: [11] Commented [17] Neerav Kingsland 8/13/2023 8:28:00 PM

https://www.aph.gov.au/Parliamentary_Business/Hansard [Author:Neerav Kingsland],[Creator:1297307460],[Date:8/13/2023 8:28:00 PM]

Page 2: [12] Commented [18] Neerav Kingsland 8/13/2023 8:32:00 PM

https://libguides.bodleian.ox.ac.uk/parliament/numbering [Author:Neerav Kingsland],[Creator:1297307460],[Date:8/13/2023 8:32:00 PM]

Page 2: [13] Commented [19] Neerav Kingsland 8/12/2023 8:52:00 PM

[redacted] [Author:Neerav Kingsland],[Creator:1297307460],[Date:8/12/2023 8:52:00 PM]

Page 2: [14] Commented [20] Nicholas Joseph 8/12/2023 9:25:00 PM

[redacted] [Author:Nicholas Joseph],[Creator:1297307460],[Date:8/12/2023 9:25:00 PM]

Page 2: [15] Commented [21] Neerav Kingsland 8/12/2023 9:30:00 PM

[redacted]
[Author:Neerav Kingsland],[Creator:1297307460],[Date:8/12/2023 9:30:00 PM]

[redacted]

