## Oxford Group on AI Policy First Week (1-3 hours)

*Please make a copy of this document via the “File” drop-down menu.*

Welcome! Over the first week, I want to determine whether, and in what ways, it would make sense to work together. To that end, I want you to complete the following tasks.

## Instructions for Submission

* Please submit your work by **11pm at your local time on** **Sunday, 5th February**.
* You’ll probably need to spend **at least 1 hour** on this week’s work.
* You probably should spend **no more than 3 hours** on this week’s work, although you are welcome to if you are particularly excited about your work. Please indicate time spent on your submission.
* Ideally, try and do this work in **one or two sittings** so that you get a smooth experience.
* Your submission should not be more than **4 pages** in the native formatting of this document, excluding these two pages of instructions.
* Please **do not include your name** in the document so that I can minimize bias on my part.
* Paste a link to your submission in the [submission form](https://docs.google.com/forms/d/e/1FAIpQLSe8MedCVrAe5tU9p8Eo-K5UJCY3eGzY-FYFNUYHiydZ1ZfSxw/viewform?usp=sf_link) (with editing rights “anyone with the link can view”). We will treat your submission confidentially.
* Please let us know if the deadline doesn’t work for you due to individual circumstances—we’re happy to accommodate.
* If you have questions, please email me, Merlin Stein, at [stein.merlin@gmail.com](mailto:stein.merlin@gmail.com)! I won’t evaluate you on your questions.

## How To Work

Let’s pay attention to:

* Clear prioritisation of tasks
* Well-structured and clear writing.
  + For example, good writing might have a sentence such as “I am unsure about X, because of 1), 2), and 3)”.
  + You don’t need to polish your writing style; just be clear.
* Results-orientedness.
* Transparent reasoning (including uncertainties).

There’s no need for full sentences or formal language. I’d prefer it if you used concise bullet points, informal language, and conceptual imaging.

## Instructions

### Project Context

*Note that more project context will be provided after the first week.*

Main Research Question: To what extent are AI Foundation Models (GPT-4 etc.) integrated into critical functions of society?

* A few specific Foundation Models are becoming more and more integrated into society; e.g., 92% of Fortune 500 use GPT-4, although it is unclear how deep this integration goes. (OpenAI [2023](https://www.youtube.com/live/U9mJuUkhUzk?feature=shared&t=121), Liang [2023](https://arxiv.org/pdf/2303.16434.pdf))
* The current literature and monitoring regime is insufficient, with no large-scale efforts.

Project Goals:

* Develop a semi-automated foundation model monitoring regime.
* Focus on the use of the models of the top 10 companies developing FMs currently (see [FMTI](https://crfm.stanford.edu/fmti/)) in (a) the cross-economy (macro picture) and (b) by companies providing critical functions (see [RAND](https://www.rand.org/content/dam/rand/pubs/working_papers/WRA200/WRA210-1/RAND_WRA210-1.pdf)).
* Monitoring on: Market shares & No. of Users. Consider B2C and B2B by sector and critical function.
  + How is it integrated? Open source vs. API vs. Finetuned partnership vs. …
  + How deeply is it integrated? Is it an essential part of the business or is it an add-on (e.g. AI chatbot/summaries for premium subscribers)?
  + How transparent is the integration? Is it clear which model is used?

Initial project approach: Descriptive base through web-scraping: Based on company registrations, databases with all company domains by sector…

### Goals for the First Week

The main goal is to get a good sense of how to advance the state of the project, focusing on 2 of the following 3 core barriers. They are sorted by priority (so if you can advance the code, I would be more excited than if you improved conceptualisation by a similar amount):

* Improving the web scraping Python code ([status quo: mini pilot](https://drive.google.com/file/d/1KbUjk9IOWCKtQsnCB6kWr4biM_HmB0cR/view?usp=sharing)) and CSV output iteratively.
* Automating the web scraping setup; e.g. writing down the key steps, automating the extraction of metadata of LLM-based chatbots to see how they were built.
* Conceptualising monitoring. Some example questions are:
  + Which search terms are the best?
    - LLMs, ChatGPT and Machine learning seem enough for now
  + What other monitoring for foundation models in critical infrastructure currently exists? For example, as part of the executive order, CISA (US Cyber Agency) receives reports on integration, and the Chinese government is making similar efforts.
    - NIST and CISA for now seem to be only monitoring agencies and I don’t see anything concrete coming from them. NIST is doing a lot of work on trustworthy ML.
  + If the CISA / UK Safety Institute / Competition Markets Authority could ask companies for specific data, what should it be about? What can they realistically provide that is helpful for these authorities and helpful for reducing systemic risks?ß
  + What should be measured with post-release monitoring outside-in and by labs?

Write up your results briefly, with an overall summary followed by clearly structured sections.

**Overall summary**

* I have designed a fairly comprehensive web scraper that is able to extract needed data to monitor the usage of foundation models across various industries.
* The current code state is not bad, but I will require a few more days to clean it up and deliver reasonable results.

The below methods sections detail the logic used in creating the system

Methods:

Company information extraction:

1. Issues:
   1. Data across both public and private companies is required to generate a meaningful look at adoption across all industry verticals. This means I needed to consult various sources to even understand a private companies market cap.
   2. Mapping a company to a specific industry vertical more a little more challenging than I expected as there are several methodologies.
2. Process:
   1. The scraper starts at the <https://www.cisa.gov/national-critical-functions-set>, here are the various big industry verticals and the corresponding sector coordinating councils <https://www.cisa.gov/resources-tools/groups/sector-coordinating-councils>. Extracting this data gives each sector coordinating council member. My scraper then searches the membership list for each council member, extracting the list of companies that are regulated by the specific federal/sector member agency. This gives us a comprehensive look at each industry.
   2. We now need to map data of each company based upon market cap. For Public and Private companies, I use the SEC EDGAR database to do a basic calculation of market cap. Gaps are further filled using CrunchBase to extract meaningful financial data. This part of the code is under development.
   3. Joining the datasets together, we now have a comprehensive look at the major players in each industry

Note: Most publicly available data I’ve seen is locked behind paywalls so I had to go to primary sources like this. I will not rule out that I have missed something as I would imagine data like this should be available quite freely, I have not been able to locate anything as of now.

AI usage per company:

Now that we have the company’s information, we need to extract information regarding their adoption of AI. For this I have identified five primary sources:

1. From the vendor company (OpenAI, Anthropic etc.) press releases/website/news claiming partnership with a customer company.
2. From the customer company claiming AI adoption into a certain product from their website/news/blogs etc
3. LinkedIn Job applications that show an uptick in particular AI related job focus
4. ArXiv papers with publications from both vendor and customer companies regarding adoption of Foundational into products
5. Github scraping of open-source repositories to see if there is an uptick in usage

Process:

* All the above methods work on a different code base but they all work on the same principal. I first use the previously obtained list of companies in order to search the company website for any usage of the term LLM/AI/Machine learning etc. If it is encountered I extract the relevant text and then summarize the data and assign it an overall topic. I am currently doing this with BART with the intention to switch to GPT4
* For ArXiv I extract all papers published by the customer company and the vendor companies and extract any mention of the company name
* For linkedin I see if there is any mention of the company and the required LLM and extract the count of entries I see

**Detailed outputs**

I don’t have a csv yet, putting all this together is taking a little more time than anticipated. I’m going to attach a sample csv and I am adding some sample output here:

From google scraping:

In 2023, Edmunds launched a pioneering ChatGPT plugin to offer users direct access to expert vehicle reviews and local inventory, enhancing informed car purchasing decisions.

(Edmunds is an automobile company)

Starting November 4, The Washington Post will introduce AI-powered audio election updates in select podcasts using Heliograf technology for personalized, localized results.

**Next steps**

If there are no concerns with the approach, then I will continue building this out.