‘Hybridity’ and the construction of Digital Futures

Work in progress
ISA World Congress, June 2023

Glen Berman, Australian National University
Kate Williams, University of Melbourne

Exploring the production of LLMs

The overarching focus of our research is ***knowledge production*** in the field of ***Artificial Intelligence*** (AI).

We draw on the concept of ***hybridity*** to investigate how AI researchers are situated within the logics of academic, commercial, and public institutions.

In the work we are presenting today, we explore, in particular, the knowledge production practices of AI researchers engaged in the development of ***Large Language Models*** (LLMs).

Hybridity in research sites

Hybrid sites are those that integrate cultural patterns, values, beliefs and practices that arise from multiple field or societal-level logics (Thornton and Ocasio, 1999).

Due to broad changes in the nature of research (variously theorised as ‘post-academic science’, ‘entrepreneurial science’, and ‘academic capitalism’), research sites have increasingly been forced to engage in commercial and media logics.

Research institutions generally have become more hybrid, and, in some instances, new research organisations have been formed for the specific purpose of better combining the distinct logics of multiple fields.

Hybridity in AI research sites

Image from “About OpenAI” (source: https://openai.com/about)

Knowledge production and LLMs

The **production of LLMs is a complex social activity**, requiring coordination across multiple fields (academic, economy, policy, media), and within these fields across multiple domains (e.g. national research funding systems and venture funding systems in the economic field).

The production of LLMs does not fit neatly into traditional field-level dichotomies (e.g. academic vs corporate) or standard domain-level dichotomies (e.g. applied vs basic research).

Rather, **LLM production occurs at hybrid sites** where these dichotomies are collapsed and renegotiated. We are focused on exploring how researchers navigate and engage in this work.

Research questions

1. On the level of individuals, who is contributing to the production of state-of-the-art LLMs?
2. How do individuals engaged in LLM knowledge production position themselves in relation to the multiple fields (academic, economic, media) within which they are situated?

Across these questions, we contrast the production of LLMs with knowledge production in the Natural Language Processing (NLP) community.

Research method

A comparative analysis of author and publication characteristics across two datasets.

Dataset 1 – LLM dataset:

1. For each LLM, identify all associated publications (academic and grey literature)
2. Extract all authors from these publications
3. For all authors, extract affiliation information and publication history. Classify authors in terms of discipline, location, institution.

Dataset 2 – NLP dataset:

1. Extract all publications from top NLP venues in the years covered by the selected LLMs.
2. For all publications, extraction author details and information.

Research method

Data analysis:

1. Patterns across LLM dataset publications:
    a. Where are they published? In peer reviewed venues? On ArXiv?
    b. What is published, and what is not? I.e. to what extent is the LLM itself made available to the public.
    c. What is the role of grey literature vs. academic literature? Are these distinctions relevant?
2. Comparison between the LLM dataset and the NLP dataset:
    a. Qualitatively, how are the same LLMs talked about across the two datasets?
    b. How are authors active across the two datasets? Do the two datasets represent the same author communities?
    c. How do sectors, fields, etc., change across the two datasets?

Areas for feedback and discussion

As we are still in the dataset development stage of this research, we welcome all feedback.

We are particularly interested in suggestions for improving our research conceptualisation and design, and in recommendations for related works or theoretical lenses to consider.