

Team

Team Members

Facilitators

Gabriel Pereira, Marc Tuters, Sabine Niederer, Carlo De Gaetano, Lonneke van der Velden, Riccardo Ventura

Participants

Ray Dolitsay, Hanna Lauvli, Lisabeth Matyash Leonsins, Mariana Fernandez Mora, David Schogt, Anvee Tara, Jiahui Xing, Angxiao Xu

Vibe Mapping

Vibe Mapping New Media Studies: Hybrid Infrastructures for Reflexive Fieldwork

Table of contents

Table of contents	1
Summary of Key Findings	2
1. Introduction	2
2. Initial Data Sets	3
2.1 The data-set	3
2.2 Data cleaning	4
2.3 Model fiddling	10
3. Research Questions	13
4. Methodology	13
5. Findings	14
5.1 Emerging Topics	15
6. Discussion	16
6.1 Comparative Insights: Complementary Perspectives	17
7. Conclusion: Abductive Infrastructures for Reflexive Fields	17
8. References	18
9. Attachments	18
Model fiddling tests	19
Baseline Test	19
Test 1 — Lowering n_neighbors to 5	19
Test 2 — Reducing n_components to 2	20
Test 3 — Increasing min_cluster_size to 12	21
Test 4 — Increasing top_n_words to 20	22

Summary of Key Findings

This project maps the evolving landscape of New Media Studies over the past 25 years by combining semantic clustering, topic modelling, and narrative synthesis—developed through a locally-run small language model. Drawing inspiration from handbooks like *Keywords for Studying Media, Culture & Information* and *Critical Terms for Media Studies*, alongside an interest in AI-assisted methods for topic modelling, we inductively constructed a field guide from 28,624 articles, identifying key conceptual formations and their changing prominence over time. We were particularly interested in mapping not just stable concepts, but dynamic thematic patterns and cultural formations—what we provisionally refer to as “vibes.”

From early concerns with technological determinism and internet history to the rising dominance of AI and platform governance, the field has continually remediated its own epistemic infrastructures. Rather than simply tracking content, we approached this as a “vibe-mapping” exercise—mapping the shifting trajectories that structure New Media scholarship through its semantic meaning.

One striking insight is the dramatic recent surge in work on artificial intelligence—especially in 2023–2024—highlighting a shifting centre of gravity in the field. Yet to focus solely on AI would risk presentism. The longer view reveals New Media Studies as a field of mediation in multiple senses: methodologically hybrid, temporally layered, and infrastructurally reflexive. Recurring threads such as surveillance & privacy, mobile media, and remix culture echo across decades, while emergent concerns—like digital memory, gender queerness, and hate speech—point to new paradigms of engagement.

The project also reflects critically on its own methods, questioning what it means to prototype scholarly tools with “small AI.” By working with locally-run LLMs for vector embedding, we seek to reimagine academic infrastructures not as extractive or commercial, but as collaborative, sustainable, and epistemically generative. This methodological commitment parallels the conceptual goal of the project: to co-develop scholarly infrastructures that are attuned to the language, logics, and values of the field itself. In doing so, Vibe-Mapping New Media contributes both a historiographic synthesis and a speculative gesture: a bottom-up field guide, shaped by the resonances that circulate through—and constitute—the field itself.

1. Introduction

This project set out to explore a dataset of publications in the field of “new media” to develop an encyclopaedia of concepts (or a book/reader of key terms or another similar interface), including references for further research—potentially resulting in a novel kind of field-defining reference work, not unlike a handbook but generated inductively via citation and semantic mapping techniques.

This work builds on earlier meta-analytical approaches to mapping scientific domains, such as co-word and co-citation analysis (Callon et al. 1986; Callon, Courtial, and Laville 1991),

applying these methods to the field of New Media in a novel, LLM-assisted context. But the work is also driven by a wish to create our own interfaces for doing research, based on the specific questions and needs of our field. What if these tools followed our values and interests? What if these tools could be locally-run, as opposed to owned and operated by large tech conglomerates? What if LLMs could be used in a way that wasn't so polluting?

Therefore, in parallel to developing a small AI tool for navigating New Media scholarship, the project also aims to reflect critically on the discourse surrounding 'small' or 'minor' AI. What imaginaries, metaphors, and problem framings guide this emerging field? What values are embedded in these alternatives—and how might they reshape our understanding of AI's role in research? These reflections inform both our design and our critique, and are elaborated in the section titled "[Small AI](#)".

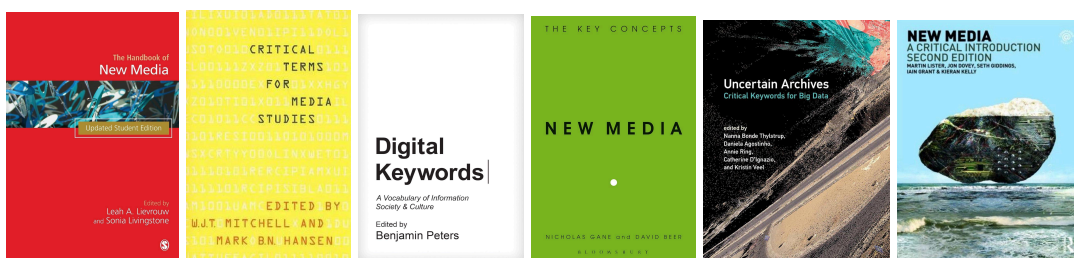
This report makes three key contributions. **First**, it introduces a novel hybrid methodology that combines expert-curated taxonomies with inductive topic modeling to surface both established and emergent conceptual structures in New Media Studies. **Second**, it offers a concrete implementation of this method using open-source, small-scale LLM infrastructure—demonstrating a viable alternative to commercial AI tools. **Third**, it reflects critically on the epistemological and infrastructural assumptions embedded in both deductive and inductive approaches, contributing to wider debates about reflexive research methods, field formation, and the politics of knowledge organization.

2. Initial Data Sets

The Vibe-Mapping New Media project explores the use of small, locally-run language models to inductively map the conceptual structure of New Media Studies.

2.1 The data-set

Baseline Corpus Construction for the Vibe-Mapping New Media Project



We constructed a baseline corpus for New Media Studies through a multi-step process combining expert curation and LLM-assisted taxonomy generation. We extracted key concepts using handbooks, and curated online glossaries. The result was a field-specific conceptual taxonomy designed to support our vibe-oriented exploratory semantic mapping.

We scraped online lists and glossaries (e.g., Wikipedia's digital media glossaries, AI ethics vocabularies, participatory design sites) to capture current and emerging terms. Lists were collected via snowballing (eg. known glossaries, calls for glossaries and references from

those glossaries to inspirational lists) and prompting ChatGPT for glossaries of key terms in New Media and Digital Culture. After review, the following list of urls was selected

- <https://textaural.com/keywords/>
- <https://datasociety.net/library/keywords-of-the-datafied-state/>
- <https://policyreview.info/glossary/introducing-decentralised-technosocial-systems>
- <https://purdue.edu/critical-data-studies/collaborative-glossary/about-glossary.php>
- <https://ainowinstitute.org/collection/a-new-ai-lexicon/page/2>
- <https://open.oregonstate.education/new-media-futures/back-matter/glossary/>
- <https://uw.pressbooks.pub/mediacriticalapproaches/back-matter/glossary/>
- <https://drstephenrobertson.com/glossary/>
- https://folgerpedia.folger.edu/Glossary_of_digital_humanities_terms
- <https://dsg.northeastern.edu/dsg-glossary/>
- <https://digitalfuturesociety.com/glossary-page/>
- https://culturalpolitics.net/index/digital_cultures/glossary
- <https://www.bloomsbury.com/uk/posthuman-glossary-9781350030244/>

For our corpus of handbooks we selected *Keywords for Studying Media, Culture & Information*, *Critical Terms for Media Studies*, *The SAGE Handbook of Media Studies*, *Uncertain Archives*, as well as lists

- Coyer, K., & Dowmunt, T. (Eds.). (Year unknown). *The alternative media handbook*. Publisher.
- Dewdney, A., & Ride, P. (Eds.). (Year unknown). *The digital media handbook*. Publisher.
- Flew, T., Holt, J., & Thomas, J. (Eds.). (2023). *The SAGE handbook of the digital media economy*. SAGE Publications [OBJ]
- Gane, N., & Beer, D. (Eds.). (Year unknown). *The key concepts new media*. Publisher.
- Hjorth, L., & Khoo, O. (Eds.). (Year unknown). *Routledge handbook of new media in Asia*. Publisher.
- Kennerly, M., Frederick, S., & Abel, (Eds.). (Year unknown). *Information keywords*. Publisher.
- Lister, M., Dovey, J., Giddings, S., & [Incomplete]. (Year unknown). *New media: A critical introduction* (2nd ed.). Publisher.
- Lievrouw, L., & Livingstone, S. (Eds.). (2006). *Handbook of new media: Social shaping and consequences of ICTs*. Sage.
- Mitchell, W. J. T., & Hansen, M. B. N. (Eds.). (2010). *Critical terms for media studies*. University of Chicago Press.
- Ott, B. L. (Year unknown). *Critical media studies: An introduction*. Publisher.
- Ouellette, L., & Gray, J. (Eds.). (Year unknown). *Keywords for media studies*. Publisher.
- Peters, B. (Ed.). (Year unknown). *Digital keywords*. Publisher.
- Rohlinger, D. A., & Sobieraj, S. (Eds.). (Year unknown). *The Oxford handbook of digital media socio*. Publisher.
- Thylstrup, N., et al. (Eds.). (Year unknown). *Uncertain archives*. Publisher.
- Trepte, S., & Masur, P. K. (Eds.). (Year unknown). *The Routledge handbook of privacy and society*. Publisher.

- Whittaker, J. (Ed.). (Year unknown). The cyberspace handbook. Publisher.

Primary Corpus for Vibe Mapping

Our primary corpus consisted of **28,624 publication records** (titles, abstracts, metadata) drawn from **35 scholarly journals** associated with New Media Studies. These included venues from a varied set of disciplines and topics across media theory, digital culture, science & technology studies, and communication studies. We collectively decided on which journals best fit with a broad understanding of “new media studies”. A key limitation emerged here: we worked exclusively with English-language journals, thus biasing the data set towards Europe and the USA, thus reinforcing the problematic reification of English as ‘lingua franca’ (see Suzina 2021).

This dataset was retrieved using the [VibeCollector](#) tool. The tool uses different sources to gather the data: CrossRef, SemanticScholar, and OpenAlex (in this order of preference). For articles without date or abstract, a different source was used to enrich the data. Some journals could not be included because they were not indexed by the data sources or did not have an ISSN number.

Final data-set:

- The full data-set organized by article (ISSN as file name):
<https://github.com/gabrielopereira/VibeMappingDMI/tree/main/fullfetchddata>
- Data-set used in our research, spanning 2000-2024:
https://github.com/gabrielopereira/VibeMappingDMI/tree/main/dynamic_topic_results_without2025

2.2 Data cleaning

We followed a 5-step program to clean the data-set **dynamic_topic_results_without2025**

Please note: when running these scripts, make sure the file titles are updated correctly and correspond to the file title you're running it on (all scripts here:
<https://github.com/gabrielopereira/VibeMappingDMI/tree/main>)

1. Remove <jats:p> and </jats:p>
To remove the <jats:p> and similar XML-tags that journals often use, we ran the below Python script: **remove_jats_tags.py**
2. Separate Book Reviews from the main dataset
We then separated book reviews from the data-set by running **Extract_book_reviews.py** (it separates the book reviews into a separate csv file)
3. Separate Editorials from the main dataset
To split and remove editorials from the data-set we ran: **split_editorials.py** (this one also creates separate csv files).

4. Re-create full data set: 25 years of New Media (2000 - 2024). Divided year by year
After cleaning, we removed years before 2000 and the year 2025 (considering it's an ongoing year with still incomplete records).
5. Creating topic keywords for analysis
For creating the keywords for each topic, BERTopic was used to remove stopwords. We used the collection of stopwords from the CountVectorizer library (in English) and a minimum presence in 3 of the topic's documents (thus removing infrequent keywords from surfacing). We enabled longer keywords, up to 3 words long (e.g. so "algorithmic surveillance" could be a word). We also used BERTopic's c-TF-IDF function, which refined keywords according to their relative importance to each year (evolutionary) and topic (global).

(Optional) Verification step: counting objects

To count the amount of objects per data-set and thereby verify the effectivity of previously described data-cleaning (across devices), we used a python script called **count_objects.py**

2.3 Model exploration

We explored the different hyperparameters of the model to test out how they shape the outputs. For this, we used a [Google Colab](https://colab.research.google.com/) and a local Jupyter Lab (<https://jupyter.org/try-jupyter/lab/>).

Following BERTopic best practices guidelines the following hyperparameters were tested and adjusted:

```
n_neighbors=n  
n_components=n  
min_cluster_size=n  
top_n_words=n
```

Final code used to run BERTopic:

<https://github.com/gabrielopereira/VibeMappingDMI/blob/main/dynamictopic.py>

See an example of model exploration in [9. Attachments](#)

3. Research Questions

Departing from our baseline analysis, our core questions were:

- What key concepts define the field of New Media Studies today?
- Can we construct a bottom-up conceptual map using both hand-curated and LLM-assisted techniques?

- How do these methods compare to traditional top-down handbooks or expert taxonomies?

4. Methodology

Deductive

Our methodology for creating the baseline began by analysing influential handbooks and lists of keywords. From these texts, we extracted 500 conceptual terms, normalising and deduplicating them. We prompted LLMs with seed terms to generate concept expansions, particularly in thematic clusters. All terms were organised into a hierarchical structure across thematic categories (e.g., *Media Temporalities*, *Aesthetics & Experience*, *Knowledge & Epistemology*, *Posthuman & Speculative*). This resulted in an *expanded conceptual taxonomy* of the field. This mapping emphasized the conceptual vocabularies, critical orientations, and emergent sensibilities that define the reflexive and theoretical dimensions of the field. It highlighted how New Media Studies articulates concerns around power, representation, affect, and infrastructural design, often through neologisms, poetic metaphors, and speculative framings.

Inductive

To inductively “vibe map” 25 years (2000–2024) of New Media Studies, we used BERTopic semantic topic modelling with vector embeddings. Titles and abstracts were combined and embedded using the small model Snowflake/snowflake-arctic-embed-s to capture the semantic meaning of each publication. BERTopic uses UMAP for dimensionality reduction and HDBSCAN for clustering.

We tested different hyperparameters across four benchmark years (2005, 2010, 2018, 2024) to evaluate stability and conceptual shifts. The outcome was the clustering of semantically similar articles, each represented by a set of keywords, refined through stopwords removal using c-TF-IDF. Outliers were connected to their most probable topic so all articles had a topic.

BERTopic then subdivides documents in each topic by year, allowing to see the evolution of the keywords used.

After inductively mapping the field across the years, we fed all the spreadsheets for all the topics into an LLM (ChatGPT-4o). The model was prompted to: (1) extract the 10 most frequently occurring non-common terms from the title_abstract fields, (2) compress those into a short list of representative keywords, and (3) suggest descriptive titles for each topic.

topic	top_keywords	compressed_keywords	final_matched_concept
0	['ai', 'human', 'artificial', 'intelligence', 'social', 'systems', 'algorithmic', 'data', 'algorithms', 'machine']	ai human artificial intelligence	Artificial Intelligence
1	['game', 'games', 'play', 'video', 'gaming', 'players', 'player', 'media', 'digital', 'new']	game games play video gaming	Game Studies
2	['television', 'media', 'article', 'tv', 'public', 'new', 'cultural', 'service', 'series', 'social']	television media article tv	Television & Public Media Cultures

3	['surveillance', 'social', 'paper', 'data', 'article', 'cctv', 'studies', 'public', 'new', 'technologies']	surveillance social paper data	Surveillance & Data Governance
---	--	--------------------------------	--------------------------------

We then verified all the topic titles manually guided by domain expertise and the deductive field mapping. We found the topics, in general, were coherent. However, we removed 8 topics that were unrelated to new media (e.g. corrigendums). In another step we fed a spreadsheet with all the title_abstracts per individual topic into the LLM. In many cases this seemed to provide good labels that we fine-tuned manually.

Building upon this initial baseline research, we then analyzed the relative frequency of publications per topic over the past 25 years. This temporal analysis allowed us to identify key shifts, continuities, and emerging patterns within the field, and informed a series of interpretive “vibe stories” that narrativize these conceptual developments.

5. Findings

Baseline

The final taxonomy comprises two interlocking layers: one deductive and one inductive. The deductive layer includes approximately 500 discrete entities extracted from over 30 canonical handbooks in the field. These concepts—ranging from “interface politics” to “cultural techniques”—provided a structured baseline from which we could evaluate the contours of the field. Together, these layers offer a dynamic and situated vocabulary for the field—attuned both to institutional memory and to evolving cultural and technological conditions.

Algorithm: algorithm, algorithmic bias, algorithmic governance, algorithmic moralism, algorithmic racism, algo-informatic bodies, automated suspicion, computational opacity, computational parallax, content moderation politics, machine unlearning, **Archive:** annotated archival description (ead), archive, archival data, haunted archives, metadata, descriptive metadata, digital editions, digital exhibits, digital object, digital preservation, digitextuality, file formats, handle, institutional repository (ir), permanent url, **Data:** big data, big data swindling, data, data broker, data cleaning, data colonialism, data curation, data extraction, data hauntology, data management, data mining, data modeling, data publics, data profiling, data sovereignty, data visualization, database, dataveillance, missing data, raw data, **Digital Identity:** anonymity, default identity, digital ids, digital identity, identity, identity infrastructures, identity tourism, name disambiguation, profile, synthetic subjectivity, user, user-centric technology, user complicity, **Infrastructure:** affordance, affordances of social media, api, application programming interface (api), apparatus, apparatgeist, architecture, cascading style sheets (css), command-line interface (cli), content management system, controlled vocabulary, digital repository, digital scholarship commons (dsc), interface, infrastructure, institutions, repository, sensorial infrastructure, software, structured data, style sheet, interoperability, network protocols, **Platform:** aca-fan, affinity portals, api, application programming interface (api), blog, blogebrity, counterpublic platforms, platform, platform capitalism, platform governance, platform realism, social networking sites (snss), user interface, comment sections, alt-tech platforms, **Surveillance:** algorithmic governmentality, ambient authoritarianism, dataveillance, disinhibition, epistemic injustice, gendered interfaces, identity infrastructures, intersectional surveillance, panoptic, privacy by design, privacy enhancing technologies, surveillance, surveillance capitalism, technopolitical imaginaries, predictive policing, behavioral tracking, **Affect & Embodiment:** affect, affective arrangement, affective economy, affective infrastructures, affective networks, affective publics, affective turn, aesthetics, authenticity, embodiment, empathy, glitch feminism, haunted media, mediated intimacy, posthuman bodies, queer temporality, sensorial infrastructure, situated knowledge, symbolic interactionism, mood, vibes, **Politics & Governance:** activism, bias, carcerality, censorship, civil law, critical legal studies, cultural imperialism, data justice, disembodiment, dissent, epistemic injustice, exclusion, freedom of speech, hegemony, human rights, indigenous data sovereignty, liberal democracy, political ecology, power, precarity, privilege, public administration, public interest, public interest technology, reparative media,

resonant misinformation, surveillance capitalism, tactical media, technopolitical imaginaries, toxic technoculture, authoritarian populism, infrastructural critique, **Cultural & Media Theory**: (in)justice, aca-fan, art, art in the anthropocene, audience, audience commodity, audience fragmentation, blog, community, cosmopolitan, cosmopolitics, cultural studies, infotainment, media effects, media literacy, mediation, memes, myths, narrative, participatory culture, passive audience, phenomenology, pleasures, plot, posthumanism, prosumer, scopic regimes, semiotics, slow media, smart, social good, story, structuralism, virtual community, virtuality, cultural analytics, media materialism, **Technology & Design**: 3d models, analog, analog nostalgia, analog technology, application programming interface (api), artificial intelligence, artificial scarcity, blockchain scenes, cascading style sheets (css), command-line interface (cli), computational opacity, computational parallax, content management system, controlled vocabulary, data visualization, digital publishing, empathic computing, game gaze, gamification, hardware-software-wetware, interface anxiety, interoperability, json, leet, machine learning, materiality, platform realism, speculative infrastructures, speculative visualization, technological determinism, technological mysticism, technology for social inclusion, text analysis, text encoding, text mining, topic modeling, virtual reality, visualization, wearable computers, xml, user-centered design, HCI, UX/UI, **Media Temporalities**: queer temporality, slow media, digital exhaustion, viral, real-time, latency, liveness, archival time, **Aesthetics & Experience**: glitch feminism, ambient noise, haunted media, post-truth affect, mediated intimacy, immersive aesthetics, techno-sublime, **Knowledge & Epistemology**: epistemic injustice, opacity politics, unthought media, post-truth affect, information disorder, platform epistemology, computational hermeneutics, **Posthuman & Speculative**: machinic desire, posthuman bodies, synthetic subjectivity, cosmic pessimism, technological mysticism, human-machine symbiosis, speculative realism, neurodiverse design, **Platform Politics & Governance**: memetic warfare, content moderation politics, user complicity, resonant misinformation, disinhibition, comment moderation, algorithmic nudging, **Digital Economies**: platform capitalism, surveillance capitalism, data colonialism, precarity, monetisation, gig economy, digital labor, creator economy, **Security & Conflict**: conspiracy, ambient authoritarianism, toxic technoculture, automated suspicion, digital conflict, cyberwarfare, information operations, adversarial media, trolling, polarization dynamics

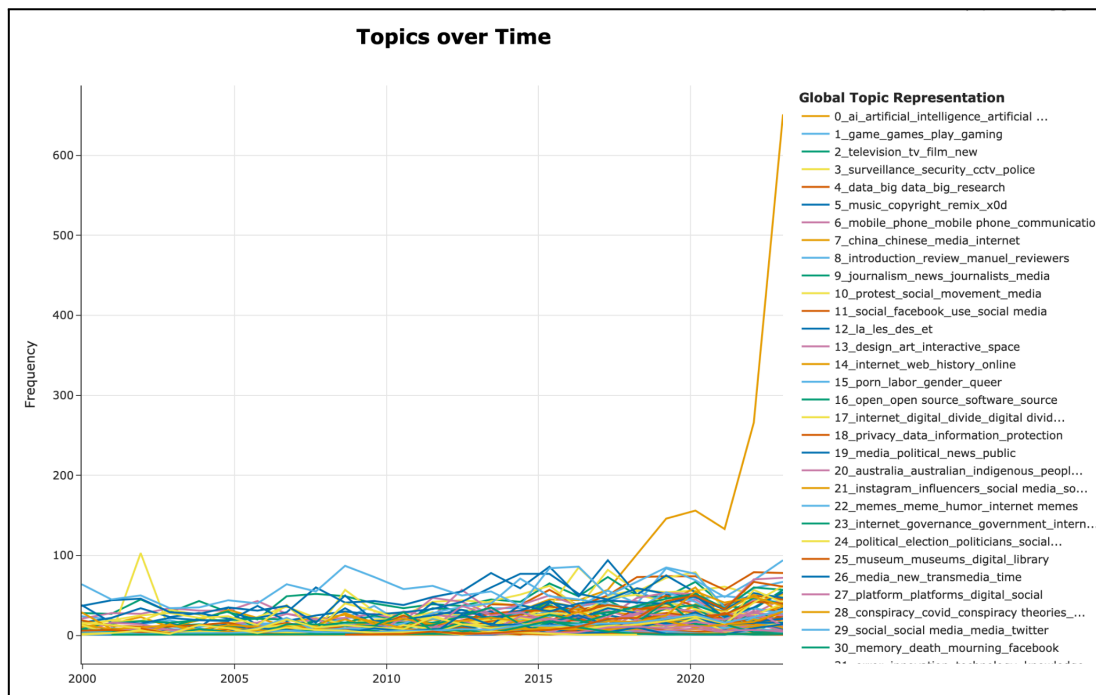
Inductive

The inductive layer was derived from the topic model trained on a curated corpus of 28,624 articles spanning 25 years of New Media Studies (2000-2024). This method surfaced latent semantic clusters that revealed both enduring research themes and surprising emergent concerns. The vocabulary uncovered here shows internal coherence—demonstrating clear associations among the titles and abstracts with keywords such as "memes," "right," and "online," or between "influencers," "Instagram," and "selfie"—as well as distinctive temporal patterns. Several emergent fields included concepts absent from canonical works, such as "quantified self," "TikTok cultures," and "toxic technoculture," validating the relevance of our hybrid expert-machine approach. The inductive approach resulted in 75 topics, but after our analysis, we decided to remove some that related to irrelevant clusters (e.g. "Editorials" or "Correction" were grouped in their own topics, indicating some still survived our initial data cleaning). Ultimately, we reached 67 topics that were validated through manual analysis of the articles present in them. These revealed a landscape composed of programmatic clusters around topics of matter for the field. In order of prominence, those topics were labelled as:

Artificial Intelligence, Game Studies, Television & Public Media Cultures, Surveillance & Data Governance, Big Data, Mobile Media & Digital Music, Mobile & Locative Media, Chinese Media Studies, News Personalization & Algorithms, Social Media Protest, Digital Aesthetics & Media Forms, Performance & Identity, Internet History, Gender, Queerness & Digital Labor, Open Source Movement, Digital Divides, Data Privacy, Media Power & Political Communication, Influencer Studies, Meme Studies, Internet Governance & Political Control, Civic Participation & Online Politics, Digital Archives, Civic Participation & Online Politics, Platform Politics & Power, Information Operations, Performing the Self: Identity, Interaction, and Expression on Social Media, Digital Memory & Commemoration, Digital Youth & Civic Engagement, Evolving News Habits and Media Trust, Feminist Technoscience Studies, Digital Learning & Higher Education, Peer Production & Online Knowledge, Misinformation & Fact-Checking, Platform Governance, Cultural & Media Theory, Information Operations, Arab Spring, Advertising, Dating, Digital Archives, Security & Conflict, Content Moderation Politics, Authoritarian Populism, Global Media Orders, Digital Ids, Born Digital, New/Digital Media, Conditions Of Possibility, Virtual

Reality, Technological Determinism, Mainstreaming Of Porn, Public Interest Technology, Communication, Biometrics & Facial Recognition, Behavioral Tracking, Data Curation, Mainstreaming Of Porn, Hate Speech, Data Curation, Toxic Technoculture, Digital Pandemic Studies, Design & Interaction, Interface Anxiety, Digital Feminism

These topics often reflected **areas of concentrated publication activity, regional or national research priorities**, and the **institutionalization of specific subfields**. They were more descriptive than conceptual, often clustered after recognizably academic research programs or methodological foci.



Caption: For our analysis, we used an interactive Plotly graph generated by BERTopic. This allowed us to manually compare and investigate different topics and their trends over time, as well as see how keywords evolved over each year. We

2008	Digital Gaming Upgrade and Recovery: Enrolling Introduction	The tagline for the 2008 Game On
2008	Seeking Truth in Video Game Ratings: Content C	Recent years have seen the video game indust
2008	The realities of virtual play: video games and the	emphasizing their degree of agency. Our analy
2008	Video games, emotion and the six senses	
2008	'You Play Like a Girl!'	Many reasons have been suggested for the lac
2008	Video Game Appropriation through Modification	This article investigates an instance in conver
2008	Review Article: The paratextual pleasures of rea	Ian Bogost, Persuasive Games: The Expressive
2008	The business of playing games: players as develo	This paper questions how video game players c
2008	Designs for learning and ludic engagement	Theories of learning and theories of play have t
2008	Adult play, psychology and design	Within psychology, play is studied mainly as a
2008	Cooperation Game	Cooperation Game is a computer-mediated s
2009	Digital Games Distribution: The Presence of the	A common criticism of the rhythm video game
2009	Save the Videogame! The National Videogame Ar	Introduction In October 2008, the UK,Â&os Na
2009	Onward Through the Fog: Computer Game Colle	In Mardi and a Voyage Thither, novelist Herm
2009	The Ecology of Games: Connecting Youth, Games and Learning	
2009	Rationalizing Play: A Critical Theory of Digital Ga	This article constructs a new framework for th
2009	From Complicity to Interactivity: Theories of Fer	Abstract This article explores the places sub
2009	Game feature and expertise effects on experienced richness, control and engagement in game	
2009	Videogame as Media Practice	Our aim in this article is to explore videogame
2009	Claiming a Stake in the Videogame	This article explores the rationalizations and
2009	Towards the Preservation of Local Computer Ga	/ New Zealand's digital game history includes
2009	Wii has never been modern: 'active' video games	This article considers the role of 'active' vide
2009	The virtual census: representations of gender, ra	A large-scale content analysis of characters i

Caption: To support in developing the stories, we manually investigated some of the titles and abstracts of articles using spreadsheets with the publications of each topic over the years. This enabled grounding our analysis and identifying some inconsistencies.

5.1 Some stories surfaced from the topics

Artificial Intelligence

Topic 0

Artificial Intelligence emerged in early discourse as a speculative and philosophical concern, tied to questions of agency, ethics, and the “conditions of possibility” for artificial life. Interest remained modest until a marked resurgence around 2020. From 2021 onward, article volume surged—peaking dramatically in 2024—reflecting AI’s mainstreaming and intensified critical scrutiny. The focus shifted from foundational questions to urgent debates around power, regulation, accountability, and machine autonomy. This arc marks AI’s journey from fringe theory to infrastructural reality, prompting renewed ethical, political, and design-oriented inquiry across the media and communication field.

Open-Source and Digital Commons

Topics 16, 44

Between 2010 and 2015, research on open-source software and digital libraries imagined the Internet as a cultural and technological commons. Scholars emphasized free software, collaborative development, public access, and the preservation of archival memory. By the late 2010s, these topics declined. Did openness, decentralization, and participation, once central to the digital future, become a background infrastructure or residual imagery in a platform-dominated present? This shift in scholarly attention may reflect a deeper reorientation in digital culture itself.

Platforms, Social Media, and the Rise of TikTok

Topics 11, 27, 29, 38

As openness declined, platforms emerged as engines of visibility, optimization, and control. Early research on social media, e.g. related to Facebook (11) and Twitter (29), focused on personal networks and public discourse. By the mid-2010s, scholars turned to platforms as governance systems (Topic 27). TikTok's rise (Topic 38) marked a shift toward short-form video, performance-based visibility, and affective labour. Across these trajectories, we see a temporal and conceptual shift: from relational connection to platforms and from networks of people to streams of content. Being online increasingly means being seen, curated, and calculated.

Surveillance and Privacy

Topics 3, 60

In the early aughts, a lot of the debate on Surveillance & Privacy focused on CCTV. In 2014 and 2015, the discussion began to focus on big data, Snowden's revelations, and other issues around digital data. In more recent years, Biometrics & Facial Recognition discussions emerged, signalling the rise of new forms of control through technological systems. The journal *Surveillance & Society* plays a key part in the debates of this topic. This also explains the blip in the year 2002: at first, we thought this could be due to the post 9/11 debates on surveillance; however upon a closer look, it seems the metadata wrongly dates articles published before 2004 as if they were all on Sep 1, 2002.

Conspiracy, Misinformation, and COVID-19

Topics 37, 28, 70, 65

Beginning in mid-2015, there was a notable rise in articles discussing bot activity and fake news on social media, especially Facebook and Twitter. This marked a shift from viewing the internet as a space of truth to recognizing its limitations and the importance of digital literacy. Between 2016–2020, articles increasingly addressed misinformation, often tied to health and vaccine concerns. The COVID-19 pandemic accelerated this trend, with a surge in discussions around conspiracy theories and disinformation spread by bots or rhetorical actors. Terms like “infodemic,” “conspiracy theory,” and “fake news” became central in New Media Studies through to 2024.

Non-Western New Media

Topics 41, 7, 39

Over the past 25 years, New Media Studies has increasingly turned its attention to non-Western contexts, with a growing body of scholarship examining developments in China, Korea, the Arab world, and other regions. A recurring theme throughout this work is the ongoing negotiation between public participation and state control. In China, media studies have evolved alongside the nation's emergence as a global power, shifting from an early focus on BBS forums and censorship to more recent explorations of digital nationalism. In the Arab world, research has centered on satellite networks and the formation of contested publics. Korean media scholarship has charted the rise of soft power through fandom and the global spread of K-pop.

Data Tracking and Digital Connection

Topics 43, 61

The trends of self-tracking and dating apps reveal how intimacy has increasingly become a core part of digital platforms, as they merge personal data with social connection. Even before 2020, location tracking and algorithmic recommendations became central to how apps operate, shaping how people connect and navigate online spaces. Self-tracking technologies, from health apps to geolocation, quietly reshaped daily life and social interactions. The spike around 2020 reflects how the pandemic accelerated these dynamics: governments normalized large-scale data tracking and lockdowns confined social interaction to digital platforms. Together, these trends show how personal data and online connections have become inseparable in everyday life.

6. Discussion

Our hybrid methodology—pairing expert-coded deductive categories with inductively derived topic models—offered a powerful way to surface the evolving vocabularies of New Media Studies. This approach moved beyond static, canon-based typologies and opened up space for:

6.1. The inclusion of emergent “vibe” oriented meanings of scholarship.

This hybrid infrastructure does not only track trends—it invites scholars to reimagine how field boundaries are constructed, debated, and rearticulated. It allowed us to move beyond static, edited collections toward a dynamic, inductively generated vocabulary for New Media Studies. Our method offered: 1) Greater inclusion of emergent and culturally inflected terms (e.g., “throbber,” “toxic technoculture”); and 2) A field-specific vocabulary useful for further semantic modeling

The comparison between the deductive (baseline) and inductive (vector embedding) mappings revealed several important patterns:

- **Thematic Divergence:** The deductive taxonomy emphasized critical theory, media aesthetics, and conceptual abstraction, whereas the inductive model surfaced empirical research clusters, often grounded in social science methodologies or policy relevance.
- **Lexical Variation:** Deductive keywords leaned toward neologisms and poetic or politicized language (e.g., data hauntology, glitch feminism), while inductive topics were more formal or field-label-like (e.g., “platform governance,” “online youth engagement”).
- **Coverage Gaps:** Concepts like reparative media, opacity politics, machinic enunciation, and queer temporality were richly represented in the deductive taxonomy but did not appear as distinct inductive clusters—likely due to their lower frequency, semantic dispersion, or non-standard terminology in academic abstracts.
- **Conversely,** inductively salient clusters such as Chinese internet studies, digital public health, and influencer marketing were underrepresented in the deductive

mapping, highlighting areas where research production has outpaced conceptual consolidation.

- **Levels of Abstraction:** The deductive approach offered a meta-theoretical lens on the field, suitable for constructing a handbook or reader. The inductive mapping provided a bottom-up view of where institutional energy is currently focused, useful for understanding disciplinary momentum and identifying emergent specializations.

6.2. Reflexive insights into the assumptions embedded in both handbooks and machine learning tools.

This study made us aware of how both canonical handbooks and machine learning models carry implicit assumptions about what constitutes knowledge, relevance, and structure. Handbooks typically assume a top-down, curatorial authority, highlighting certain concepts as foundational while marginalising others. These editorial decisions are shaped by disciplinary traditions, institutional hierarchies, and geopolitical-linguistic constraints on what gets named, circulated, and legitimised.

In contrast, machine learning tools tend to prioritize semantic coherence, statistical significance, and linguistic dispersion. However, this supposed objectivity encodes biases through training corpora, vector representations, and optimization goals. Our approach aimed to surface these epistemological asymmetries, encouraging scholars to question what is excluded, flattened, or overemphasized in both human- and machine-curated knowledge systems.

6.3. A testbed for non-commercial LLM-assisted knowledge mapping, tailored to scholarly domains.

The project illuminated how we may critically engage with alternative methods to both commercial LLMs and top-down scholarly canons. It offered a sandbox for experimenting with non-commercial, reflexive models that foreground scholarly priorities while also surfacing how algorithmic assumptions shape the very contours of legibility and relevance.

There are several points of departure that could be developed further—such as applying the model to systematic or scoping literature reviews. There could be several different entry points into such an analysis, e.g. dynamic topic modelling (timeline), topological graphs, temporal clustering and more. Moreover, the results should also inspire other ways of finding and browsing scholarly work, aligned with academic priorities rather than those of Big Tech companies—see e.g. the [VibeSearch](#) tool that inspired this project.

Finally, this project was performed on an English language corpora, which must therefore be accounted for when analysing the conceptualization of the field. This introduces cultural and epistemological limitations. Further testing using different linguistic and regional datasets should be part of the next step of analysing the models variability and addressing potential cultural or epistemological blind spots.

7. Conclusion: Abductive Infrastructures for Reflexive Fields

The taxonomy and topic model presented here establish a semantic infrastructure—for example, a structured vocabulary and latent topic map—that is, for subsequent phases of this project. Together, they enable vector-based inductive mapping and methodological experimentation with small LLMs in the humanities.

More than a static map, the hybrid model functions as a generative epistemic interface. The deductive taxonomy offers conceptual depth, while the inductive model foregrounds patterns of scholarly activity. Their combination supports what Piece called [“abduction”](#)—a mode of reasoning that begins not with proof, but with plausible inference, speculation, and hypothesis generation. Abduction is particularly suited to exploratory research, where patterns must be discerned in messy or emergent terrain.

By designing tools that facilitate abductive engagement with the field, this project reframes LLMs not as black-box summarizers, but as catalysts for collective sensemaking. This approach empowers scholarly communities to define their own epistemic terrain—not only descriptively, but reflexively and propositionally, shaping the very categories and methods through which knowledge is organized, contested, and renewed.

The hybrid model thus serves as both mirror and probe: reflecting the field’s evolving contours while also enabling new ways of navigating, inhabiting, and reimagining them. It affirms the value of experimental infrastructures that are attuned to scholarly values, semantically grounded in disciplinary practice, and open to iterative co-construction—rather than extractive or externally imposed.

Looking ahead, this abductive infrastructure may serve not only as a diagnostic map of an existing field but also as a kind of exploratory means for shaping new research agendas. By enabling field-specific vocabularies to emerge from within the community rather than from external platforms or editorial gatekeeping, this approach suggests a new kind of model, beyond bibliometrics. Future work could integrate multilingual corpora, explore more dynamic interfaces (e.g. knowledge graphs or exploratory dashboards), and apply similar methods to other interdisciplinary domains where conceptual boundaries are in flux. In this sense, vibe mapping is not just a method, but a proposal for how we might attune scholarly tools to the epistemic rhythms of reflexive fields.

8. References

- Callon, M., Courtial, J.-P., & Laville, F. (1991). *Co-word analysis as a tool for describing the network of interactions between basic and technological research*.
- Andersen, C. U., & Cox, G. (2023). *Toward a Minor Tech*.
- Munk, A. K., et al. (2024). *What Are Algorithms Doing in the Scientific Literature?*
- Suzina, Ana Cristina. 2021. “English as Lingua Franca. Or the Sterilisation of Scientific Work.” *Media, Culture & Society* 43 (1): 171–79.
<https://doi.org/10.1177/0163443720957906>.

9. Attachments

To explore how different hyperparameters affect topic modeling results, we conducted five tests using the same dataset from 2018. The first test served as a baseline with default parameter settings. In each of the following four tests, we adjusted one key parameter at a time to observe its specific impact on the output. These parameters include `n_neighbors`, `n_components`, `min_cluster_size` and `top_n_words`, which control the clustering structure, topic representation, and dimensionality reduction within the BERTopic model. This step-by-step approach allowed for a controlled comparison of how individual settings influence the number and structure of detected topics.

Model exploration

Baseline Test

Dataset: Year 2018

Hyperparameters:

`n_neighbors=20`

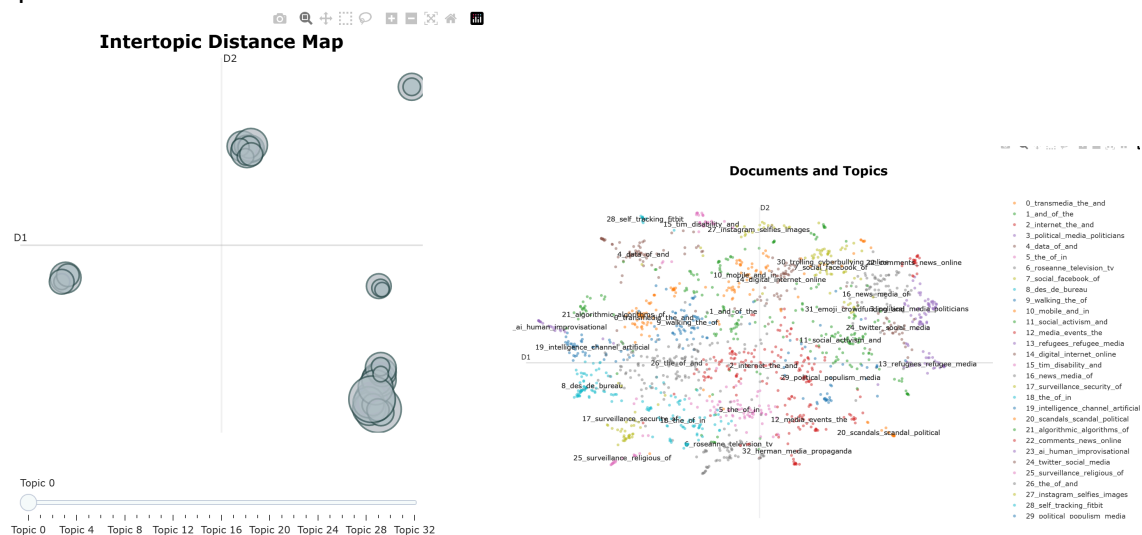
`n_components=5`

`min_cluster_size=10`

`top_n_words=11`

Output:

32 topics



Test 1 — Lowering `n_neighbors` to 5

In this test, I reduced the `n_neighbors` parameter from 20 to 5 while keeping all other settings constant. As expected, a smaller neighborhood size in the UMAP model led to more

fragmented clusters. The number of detected topics increased from 32 to 43, and the Intertopic Distance Map shows a larger overall spread with less overlap between clusters. This indicates that the model now captures more fine-grained differences between documents, but possibly at the expense of producing very small or highly specific topics. Visually, the document-topic map also shows more dispersed clusters with sharper boundaries compared to the baseline.

Dataset: Year 2018

Hyperparameters:

`n_neighbors=5`

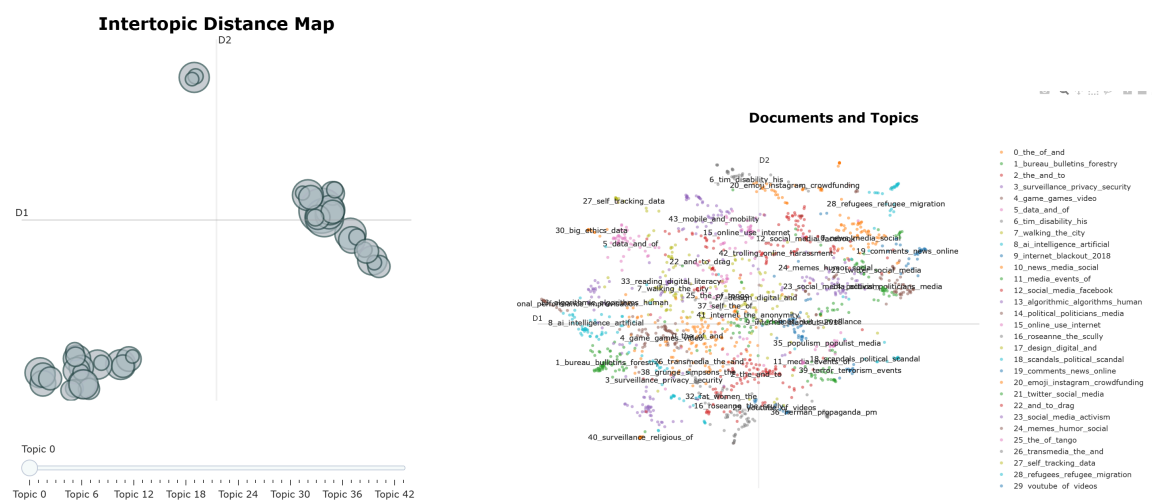
`n_components=5`

`min_cluster_size=10`

`top_n_words=11`

Output:

43 topics



Test 2 — Reducing `n_components` to 2

In this test, I reduced the `n_components` parameter from 5 to 2 to observe how lowering the dimensionality of the UMAP projection affects clustering and topic formation. The number of detected topics increased from 32 to 37, suggesting that more distinct clusters emerged. Visually, the Intertopic Distance Map shows more dispersed topic bubbles, indicating a greater separation between clusters. However, lowering the dimensionality may also exaggerate certain distances or overlaps, as some structural information is inevitably lost. The document-topic map reveals that while clusters are more visually separable, the overall distribution appears slightly noisier compared to the baseline.

Dataset: Year 2018

Hyperparameters:

`n_neighbors=20`

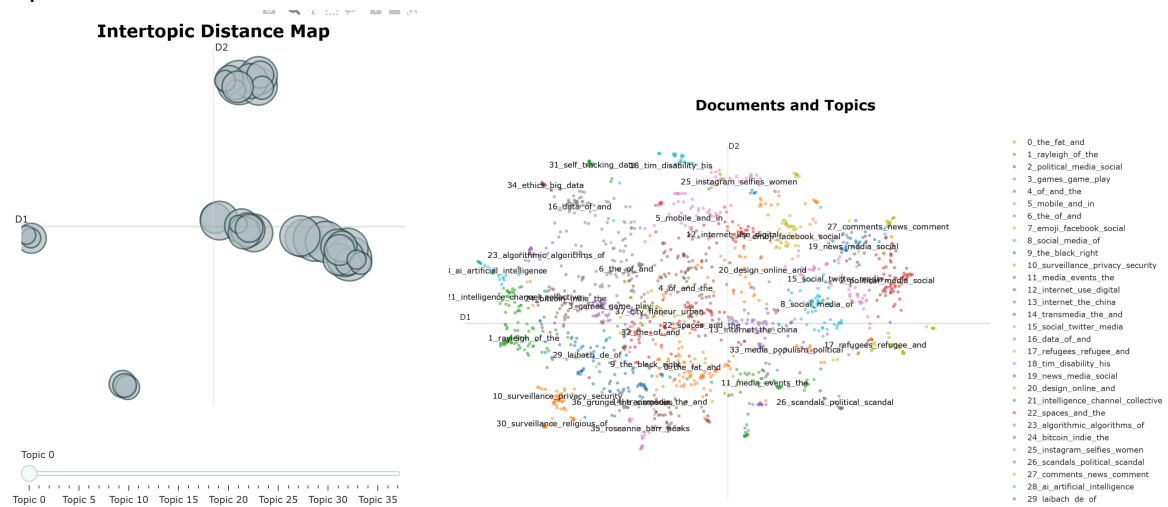
`n_components=2`

```
min_cluster_size=10
```

```
top_n_words=11
```

Output:

37 topics



Test 3 — Increasing min_cluster_size to 12

In this test, I increased the `min_cluster_size` parameter from 10 to 12 to examine the effect of stricter clustering requirements. As expected, the total number of detected topics dropped significantly from 32 to 22. The Intertopic Distance Map shows fewer, but generally larger topic bubbles, with clearer separation between clusters. The document-topic map also reflects this, with fewer scattered points and more consolidated topic groups. This suggests that a higher `min_cluster_size` reduces noise and small, overly specific topics, resulting in a cleaner but potentially less fine-grained topic structure.

Dataset: Year 2018

Hyperparameters:

```
n_neighbors=20
```

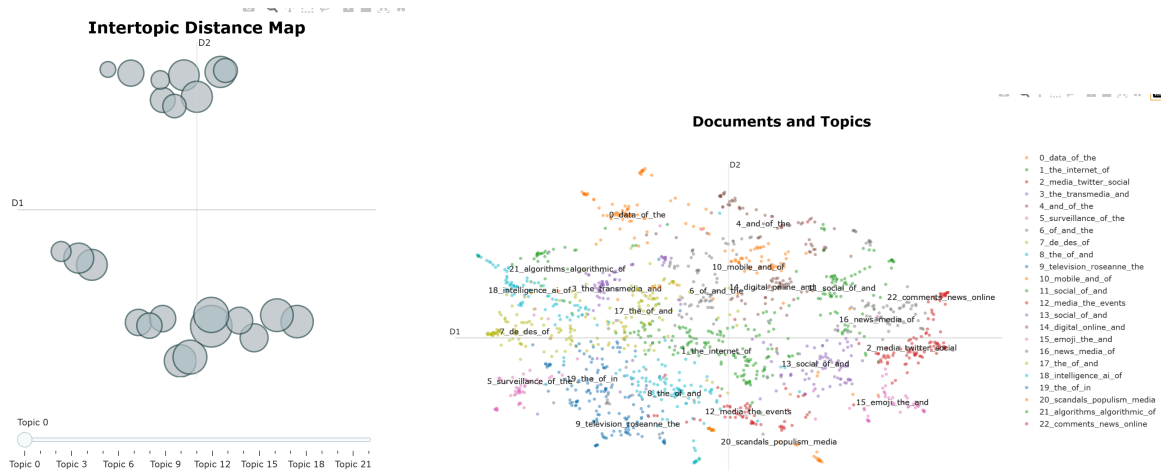
```
n_components=5
```

```
min_cluster_size=12
```

```
top_n_words=11
```

Output:

22 topics



Test 4 — Increasing top_n_words to 20

In this final test, I increased the `top_n_words` parameter from 11 to 20 to explore how expanding the number of displayed keywords affects topic structure and interpretation. As expected, this parameter change does not influence the clustering process itself, as the total number of detected topics remained similar at 33. However, the additional keywords provide more detailed topic labels, offering richer context for interpreting each cluster. Visually, both the Intertopic Distance Map and the document-topic map are consistent with previous tests, confirming that this setting mainly affects output readability rather than the underlying topic formation.

Dataset: Year 2018

Hyperparameters:

`n_neighbors=20`

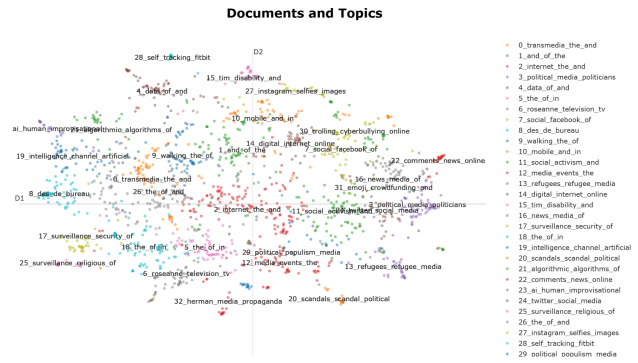
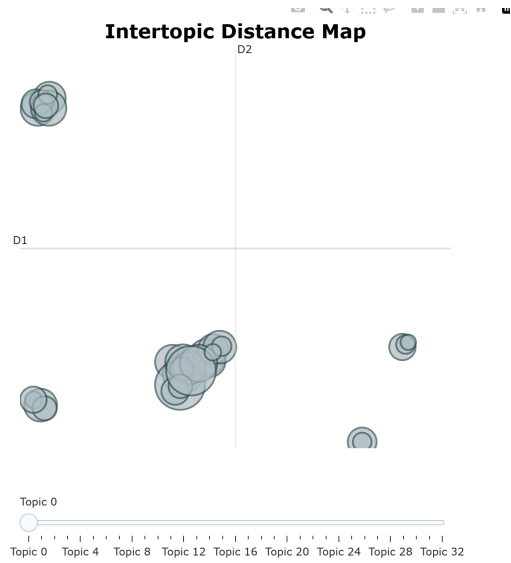
`n_components=5`

`min_cluster_size=10`

`top_n_words=20`

Output:

32 topics



Small AI

Small AI Wiki

Team Members

Facilitators

Gabriel Pereira, Marc Tuters, Sabine Niederer, Carlo De Gaetano, Lonneke van der Velden, Riccardo Ventura

Participants

Ray Dolitsay, Hanna Lauvli, Lisabeth Matyash Leonsins, Mariana Fernandez Mora, David Schogt, Anvee Tara, Jiahui Xing, Angxiao Xu

Contents

Contents	1
Summary of Key Findings	1
1. Introduction	2
2. Initial Data Sets	2
3. Research Questions	2
4. Methodology	2
5. Findings	2
6. Discussion	2
7. Conclusion	2
8. References	2

Summary of Key Findings

1. **“Small” AI as a Constellation of Values:** Smallness is not simply a technical descriptor (model size or computational resources), but a deliberate value-driven stance. Projects framed small AI around transparency, efficiency, data sovereignty, feminist ethics, ecological sustainability, and community-driven approaches, indicating a broad cultural shift toward more accountable AI.
2. **Five Distinct Value-Based Clusters:** Using network analysis, the research identified five coherent value clusters within the small AI landscape:
 - a. **Feminist and Queer Approaches** prioritize intersectionality, care, and participatory governance, building inclusive models;
 - b. **Small Data, Big Impact** emphasizes high-quality datasets, consent-driven data practices, and civic-oriented applications, challenging assumptions that larger models always outperform smaller ones;
 - c. **Open Source, Low-Compute, and High-Performance** shows how decentralization, local deployment, and open ecosystems can yield high-performing models that rival corporate systems without heavy infrastructure.
 - d. **Green AI: Low-Footprint Models for Biodiversity** demonstrates that lightweight, ethical AI models can facilitate ecological monitoring and interspecies understanding, aligning technological practice with environmental justice.
 - e. **Indigenous-led Models** center data sovereignty, cultural continuity, and community governance, positioning small AI as politically engaged and culturally responsive.
3. **Practical Translation of Values into Model Fine-Tuning:** We proposed a methodological innovation that translates abstract value frameworks into actionable, tweakable technical parameters. The research leveraged GPT-o3 and BERTopic experimentation to operationalize how different ethical or political commitments (feminist, ecological, Indigenous) concretely influence fine-tuning decisions.
4. **Participatory and Recursive Keyword Methodology:** The iterative, GPT-assisted keyword extraction and validation process successfully identified subtle yet significant distinctions in project framings, highlighting the importance of combining automated and manual methods to capture nuanced ethical and cultural meanings in AI discourses.
5. **Community and Decentralized Governance as Core to Small AI**
Across clusters, community-driven practices emerged consistently as essential elements. Whether feminist, ecological, or Indigenous, these models challenge

the extractive logics of corporate AI, emphasizing participatory governance, ethical licensing, and accountability.

1. Introduction

This research project maps “small AI” initiatives: language models and related tools that offer alternatives to dominant, large-scale generative systems. Rather than defining “small” strictly by size, we understand it as a constellation of values such as transparency, sustainability, open-source development, intersectionality, and data sovereignty.

Using a mixed-methods approach, we identified 67 initiatives through desk research, GPT-assisted deep search, and snowballing techniques. Our focus was on projects developed by startups, academics, NGOs, open-source communities, and artist collectives between 2004 and 2025. We excluded large tech companies unless their models were explicitly framed as alternatives.

From project websites and publications, we compiled keyword lists related to values and strategies. These lists were refined and expanded through iterative prompting of ChatGPT-o3, followed by manual verification to ensure contextual relevance.

We then constructed a bipartite network of projects and keywords in Gephi and used modularity analysis to identify five distinct value-clusters. These clusters were manually annotated to describe different approaches to small AI. As a final step, we translated the identified values into sets of tweakable parameters, offering a hands-on guide for experimenting with fine tuning a model in line with small AI values.

2. Initial Data Sets

The resulting dataset captures 67 distinct projects that self-identify with a “small-AI” ethos. Each entry records the project name, URL, project type, umbrella category, a description drawn from its own documentation, and a cleaned list of value and strategy keywords. Projects span **six umbrella categories**—Models & Model-Families (29 projects), Toolkits/Libraries/Platforms (14), Applications & End-User Agents (9), Communities & Initiatives (6), Research & Thought-Leadership (5), and Data & Evaluation Resources (4). Within these, 21 finer-grained project types are represented:

1. **Models & Model-Families**, including Model, Model / Artistic project, and Model series;

Projects: AfriBERTa, AI Nūshu, all-MiniLM-L6-v2, Alpaca (Stanford), BLING series (LLMware), BLOOM (BigScience), ClimateBERT, Fietje 2, FLAN-T5 Small, Gemma (2 B / 7 B), H2O-Danube-1.8B, InkubaLM (Lelapa AI), Koala (Berkeley AI), Llama 3.2-1B (secrets detection, Wiz), Magistral Small 3.1, Meerkat-7B-v1.0, Mistral 7B, NatureLM-audio, OpenELM, Orca 2 (13 B), Phi-4, Phone LM 0.5 – 1.5 B, Qwen-2 0.5B, SmolLM2, StableLM, TeenyTinyLlama, TinyLlama (1.1 B), UlizaLlama, Vicuna

2. **Communities & Initiatives**, including Feminist AI initiative, Community-led tech collective, and Community organisation;

Projects: AI + Alliance, Feminist AI, Masakhane, O'KANATA (Indigenous AI platform), Queer in AI, Te Hiku Media / Papa Reo

3. **Research & Thought-Leadership**, encompassing Manifesto, Academic paper / Manifesto, Manifesto / Artistic project, and Education;

Projects: AI Decolonial Manifesto, Data Ecofeminism (ACM FAccT 2025 paper), Data Feminism for AI (ACM FAccT 2024 paper), Little Language Models (MIT CoCo), ΑΠΟαποικιοΠΟΙΗΣΗ

4. **Toolkits, Libraries & Platforms**, including Tool / Platform, Toolkit, Training method, Infrastructure, Infrastructure library, and Research infrastructure;

Projects: ARIES – Artificial Intelligence for Environment & Sustainability, AymurAI, Divvun/Giellatekno – Sámi Language Technology Project-The Sámi Case Study (MultiLingual, Oct 2024), EDIA – Estereotipos y Discriminación en IA (Fundación Vía Libre), GPT4All (Nomic), GreenAuto, ICARUS (Animals' Early-Warning), Indigenius (CDIAL), Kōrero Māori, llama.cpp, Papa Reo (Māori-led language tech), Project CETI, SALT, Spawning AI / “Have I Been Trained?”

5. **Applications & End-User Agents**, including Chatbot and Artistic chatbot;

Projects: Botchen (Denotation UG), Maru anti-harassment chatbot, myShakti (Yotta), PocketPal AI, Queer AI, Shhor AI, SOF+IA (Chile), Tuumailubotti, Xela feminist chatbot

6. **Data & Evaluation Resources**, including Dataset / Toolkit, Benchmark, and Semantic-analysis tool.

Projects: Feminist Data Set, Projeto Sócrates (Brazil), TinyStories dataset / Phi-series, WinoQueer

[Link to spreadsheet here.](#)

3. Research Questions

RQ1: What values, principles and strategies characterize current discourses around “small AI”?

RQ2A: How do different small AI projects position their alternative approach to dominant AI?

RQ2B: How can these values be translated into tweakable parameters for fine tuning small AI models?

4. Methodology

This project maps the diverse landscape of current small AI language models and related projects. We understand the notion of “small” not simply in terms of model size or parameter count, but as a constellation of values that offer meaningful alternatives to dominant, large-scale, corporate-driven generative models. These values may include efficiency, transparency, accessibility, sustainability, feminist or decolonial approaches, and commitments to regional equity or cultural specificity.

To identify relevant projects, we used a mixed-methods approach. This included desk research using targeted queries such as “small AI language models” and “sustainable AI language models,” as well as prompting GPT-based deep search (specifically ChatGPT-4 and GPT-o3) and applying snowball sampling techniques to expand from initial seed results. We explored both formal sources (research papers, model cards, GitHub repositories) and more informal publications (blogs, manifestos, interviews, press articles).

We included initiatives developed by startups, open-source communities, academic laboratories, non-governmental organizations, and artistic research collectives. Projects associated with major technology corporations were excluded unless they explicitly framed their models in terms that aligned with the alternative values we were investigating.

For each project, we recorded the name, URL, and project type. Project types were organized under six macro-categories.

From each project’s public materials—including official websites, technical documentation, and additional news sources published between 2004 and 2025—we extracted a short description of the project’s aims, as well as direct quotes that illustrate how they frame the notion of smallness. These were particularly useful in surfacing how actors self-describe their alternative approaches in relation to dominant AI paradigms.

We used a combination of manual coding and GPT-assisted extraction (via GPT-o3) to identify values, features, and strategies mentioned by the projects. This produced a broad and sometimes overlapping keyword set. To refine this list, we prompted GPT-o3 to help normalize terms (e.g., correcting typos, merging synonyms) while preserving important nuances.

In a second iteration, we prompted GPT-o3 to revisit each project’s website and suggest additional keywords we may have missed, providing a rationale for each new inclusion. This recursive step helped expand our vocabulary and catch underrepresented framings. We verified the proposed new keywords manually in duos, discussing whether to include or exclude them from the original list. Slightly more than half of the new keywords were included.

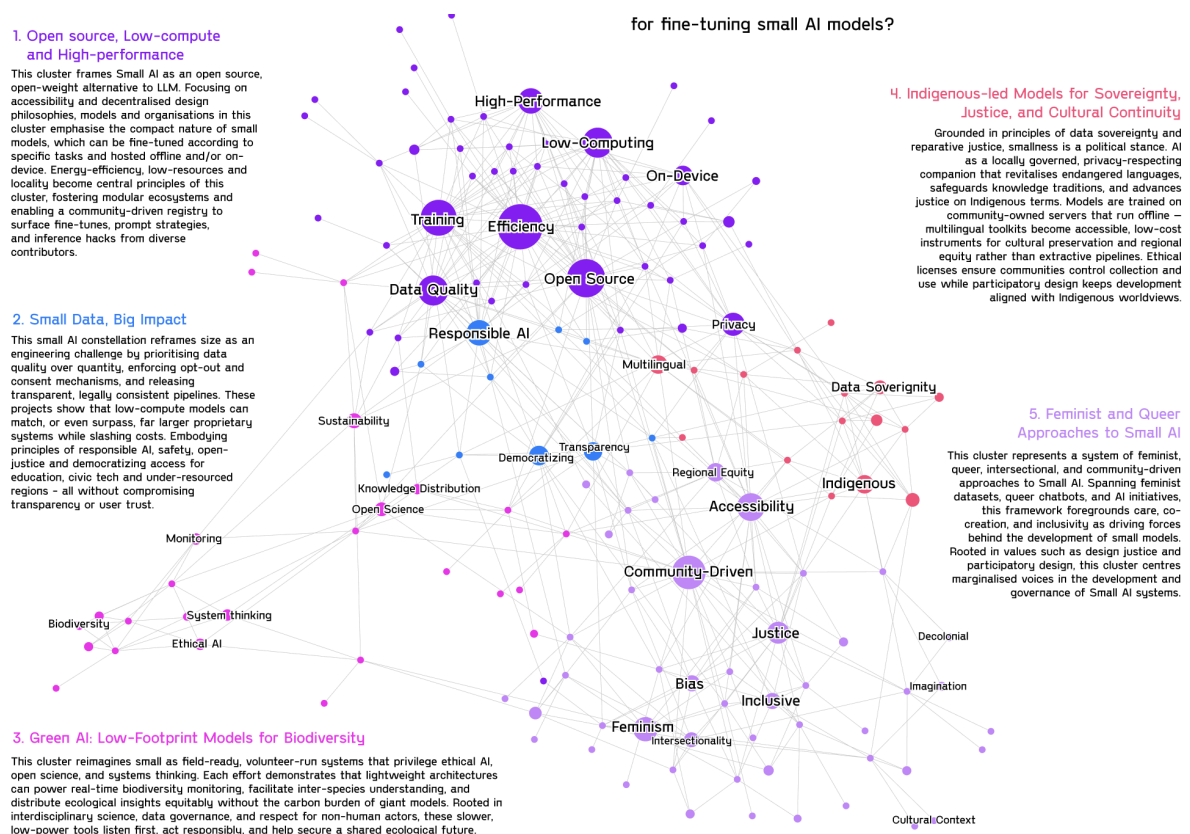
We then built a project-keyword network using Gephi and applied modularity class detection to identify clusters within the network. Five main clusters emerged, each corresponding to a group of projects that shared overlapping values or strategic orientations: Feminist and Queer Approaches to Small AI; Small Data, Big Impact; Open Source, Low-Compute and

High Performance; Green AI: Low-Footprint Models for Biodiversity; and Indigenous-led Models for Sovereignty, Justice, and Cultural Continuity.

These clusters were manually annotated to identify the underlying patterns—for instance, projects emphasizing “smallness as sustainability,” “smallness as ethics,” or “smallness as feminist infrastructure.” We then used these thematic clusters to engage ChatGPT in a structured conversation about emerging “small AI” principles.

As a final step, we prompted ChatGPT (GPT-o3) using the Google CoLab version of BERTopic, combined with the VibeCollector dataset, to explore how these principles could be operationalized. Specifically, we asked how adjusting BERTopic settings (e.g., dimensionality reduction, topic reduction thresholds) might reflect or enact the different value frameworks emerging from our analysis. What would a feminist fine-tuning of BERT look like, compared to a green one?

When “smallness” is enacted as a feminist, decolonial, or ecological stance, what new practices of care, accountability, and storytelling become possible inside the model’s very architecture? And how do these translate into tweakable parameters per value-cluster?

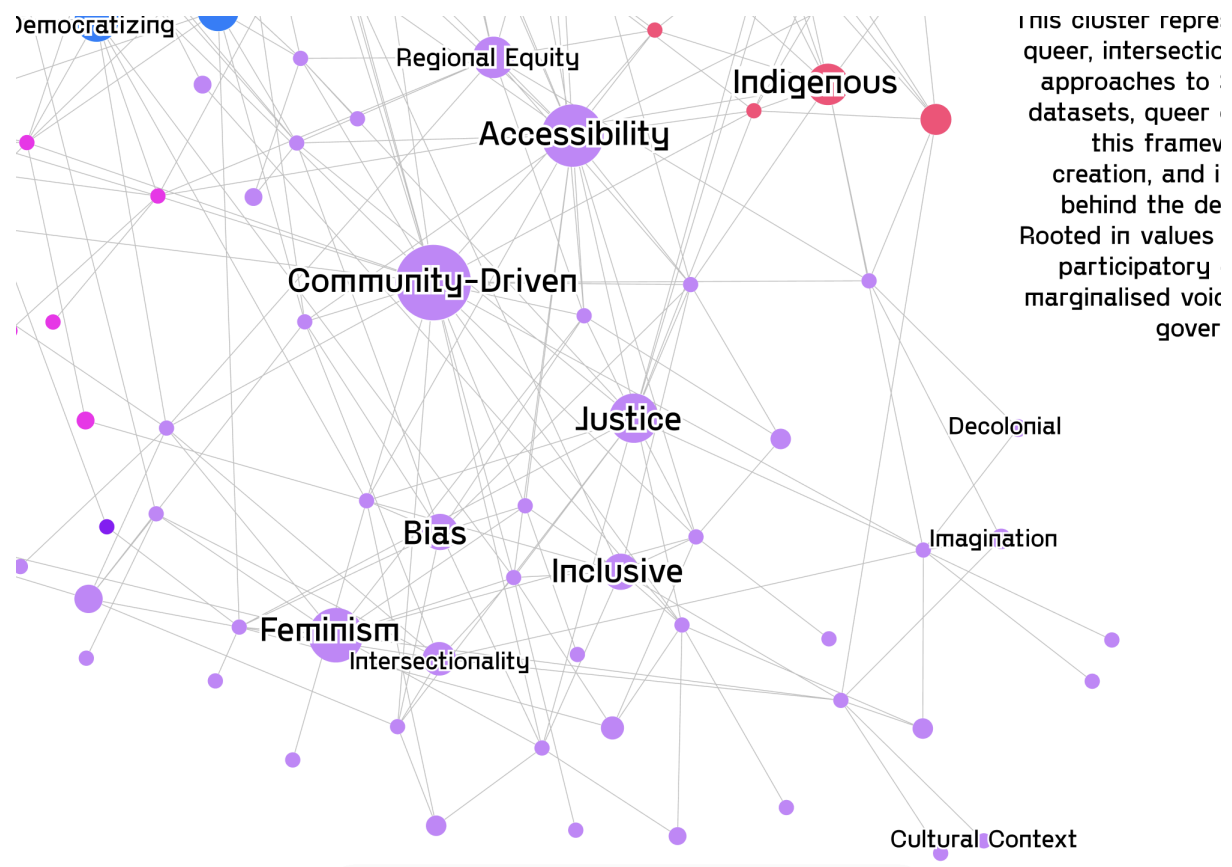


link to figma:

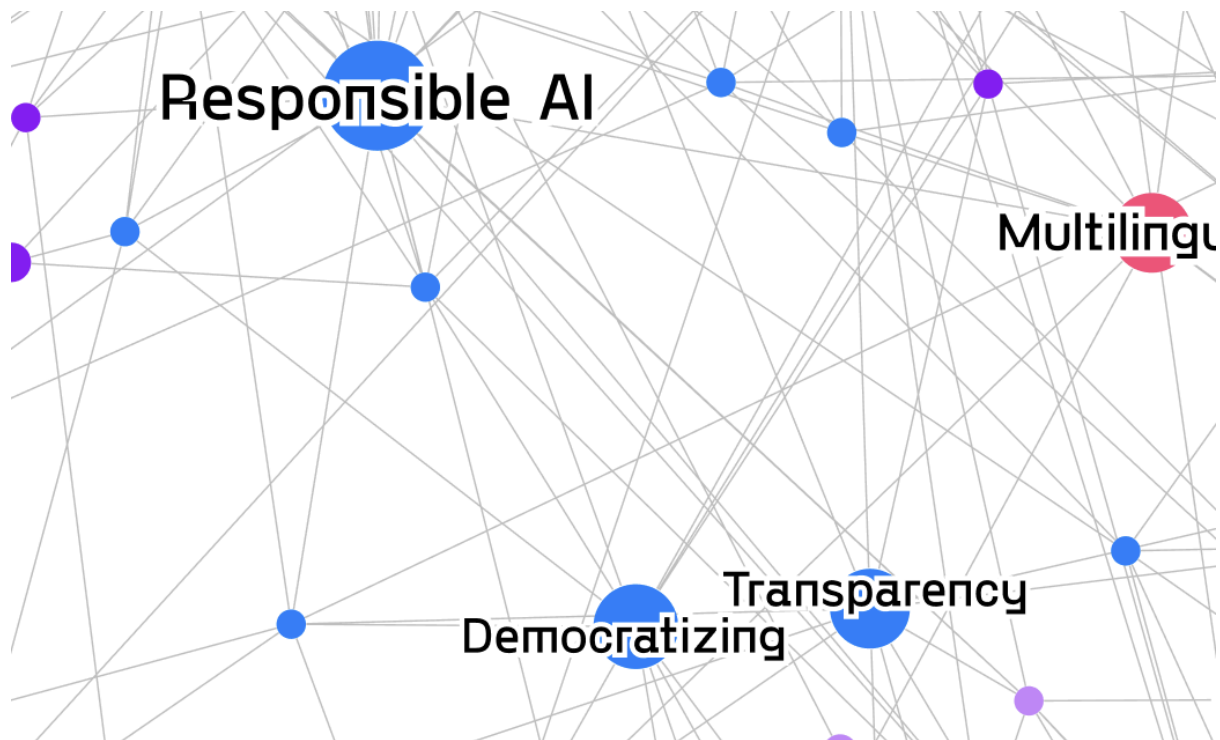
<https://www.figma.com/design/SGFXq2ecXAEUFb8Wdmaj3/VibeMapping?node-id=0-1&p=f&t=7CxcbsWfgpPJHxj-0>

5. Findings

After running the Modularity Class algorithm in Gephi, five thematic clusters were delineated.



Modularity cluster one is titled *Feminist and Queer Approaches to Small AI*. The initiatives in this cluster represent a system of feminist, queer, intersectional, and community-driven approaches to AI, including keywords such as “community-driven”, “feminism”, “inclusive”, and “queer”. These projects originate across academic, activist, and artistic spheres, including queer chatbot initiatives (Xela, Maru), feminist data infrastructures (Feminist Data Set, Data Feminism for AI), and community-aligned research outputs (WinoQueer, Little Language Models). Many projects embrace participatory and co-creative design, such as Tuumailubotti and AI Nūshu, while others interrogate bias and representation through intersectional lenses (EDIA, Queer in AI). Roughly a third of the projects center queer, feminist, or decolonial agendas; another third emphasizes participatory governance and community knowledge-sharing; the remainder explore artistic, educational, or anti-harassment applications. This framework foregrounds care, co-creation, and inclusivity as driving forces behind the development of small models. From feminist datasets and queer chatbots to ecofeminist manifestos and bias-detection tools, these projects challenge the extractivism of large models, seeking instead to build decentralized infrastructures that prioritize ethical data training. Rooted in values such as design justice, data feminism, and participatory design, this cluster centers marginalised voices in the development and governance of Small AI systems.



Modularity cluster two is titled *Small Data, Big Impact*. This small AI constellation reframes size as an engineering challenge by prioritising data quality over quantity, enforcing opt-out and consent mechanisms, and releasing transparent, legally consistent pipelines. Spanning university labs, civic-tech communities, and open-source collectives, the projects include compact instruction-tuned models (Alpaca, Koala, Phi-4, Vicuna), domain-specific deployments for under-resourced languages and publics (Aymurai, Projeto Sócrates), and infrastructural tools that guarantee user consent and dataset provenance (Salt, Spawning AI). About 40% of the initiatives centre on instruction-tuned research prototypes that push the limits of low-compute performance. Roughly a third target educational or civic-tech use cases, proving that small models can serve communities without conceding accuracy. The remaining supply governance and auditing infrastructure, foregrounding consent and transparency. These projects show that low-compute models can match, or even surpass, far larger proprietary systems while slashing costs. Embodying principles of responsible AI, safety, open-justice and democratizing access for education, civic tech and under-resourced regions - all without compromising transparency or user trust.

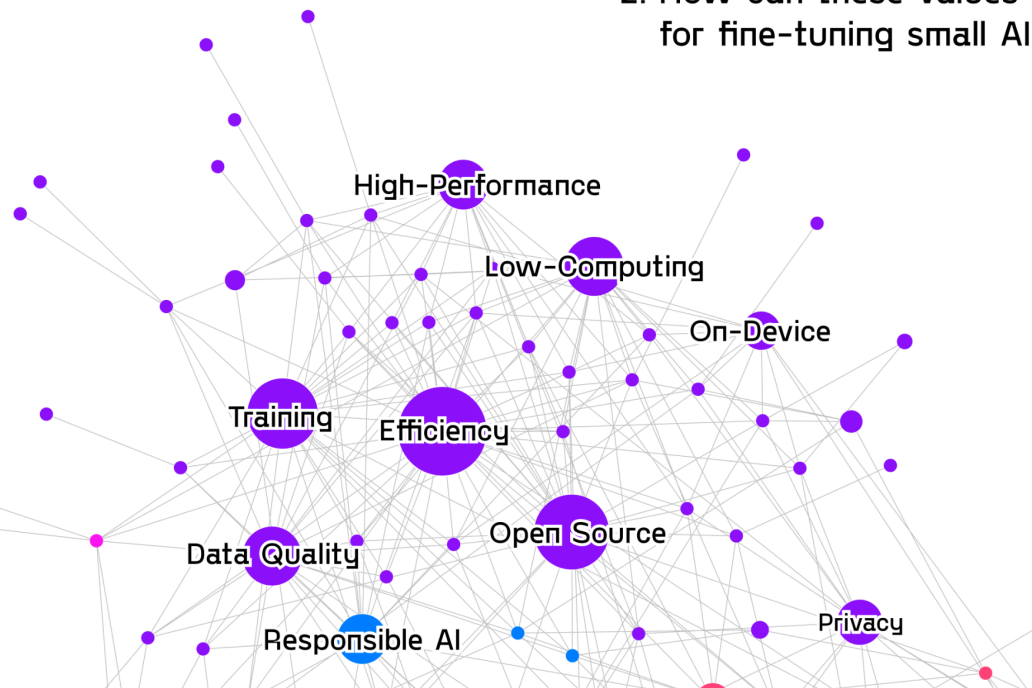
2. How can these values be realized for fine-tuning small AI

te

en source,
sing on

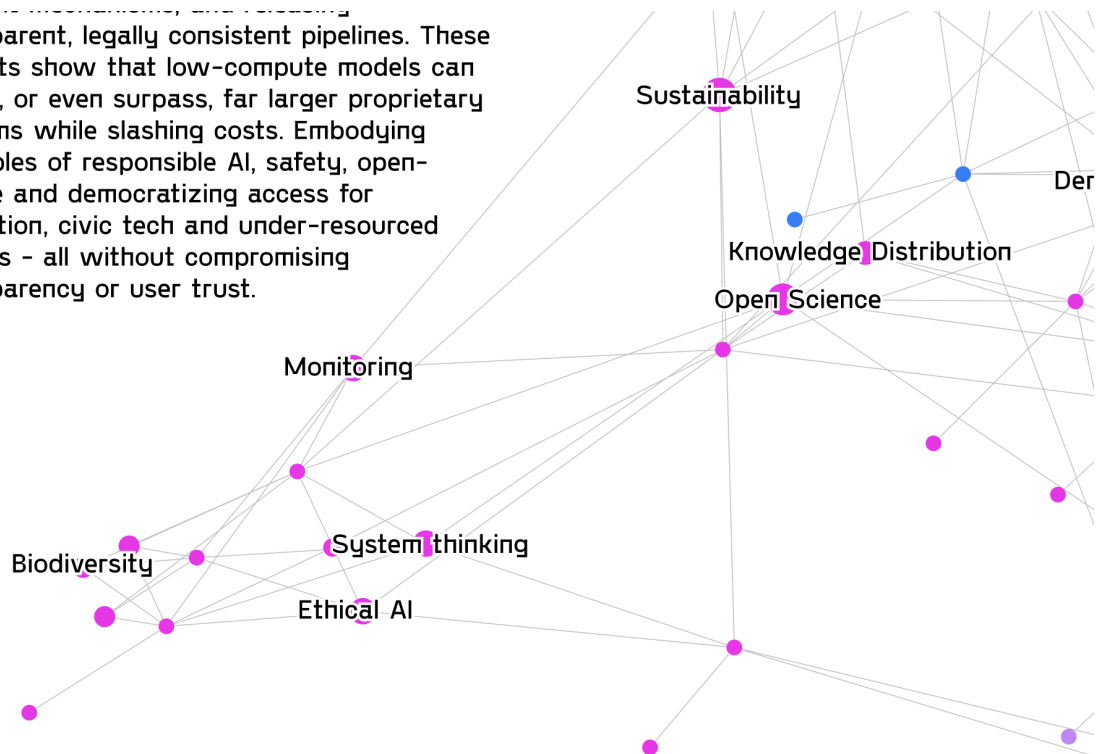
s in this
e of small
ording to
'or on-
es and
his
, and
to

ze as an
lata

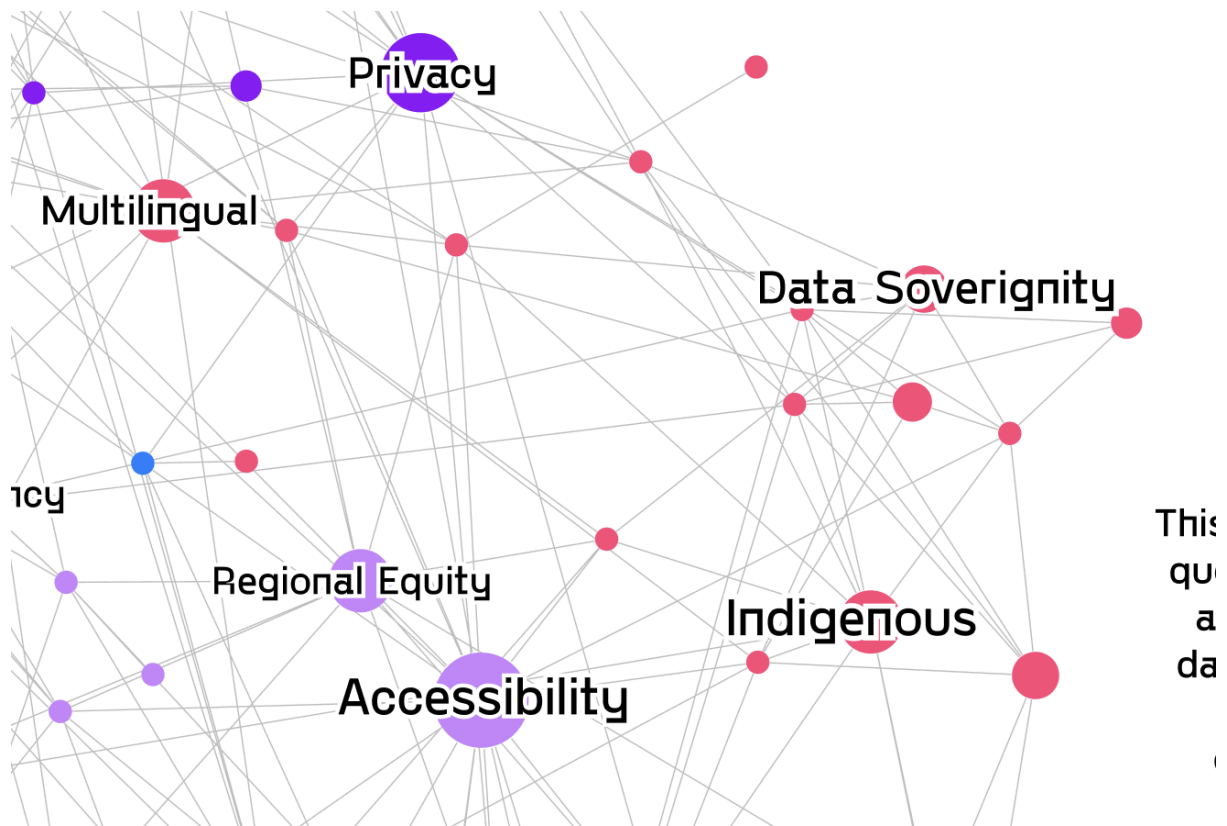


The third thematic cluster is titled *Open source, Low-compute and High-performance*. This cluster highlights a decentralized, efficiency-driven approach to Small AI, framing small models as an open source, open-weight alternative to LLMs. This cluster is organized around keywords such as “on-device,” “open ecosystem,” “low-resource,” and “energy-efficient.” Projects in this cluster include edge-ready models (TinyLlama, Phone LM, TeenyTinyLlama), sustainability-oriented LLMs (GreenAuto, ClimateBERT), and lightweight fine-tuning platforms (Gpt4All, Botchen). Many initiatives prioritize transparent documentation and reproducibility, such as OpenELM and H2O-Danube, while others focus on quantized, offline deployment using tools like llama.cpp. A significant number of projects support language or region-specific applications (e.g., AfriBERTa, Fietje 2, Inkubalm); while others emphasize privacy-preserving, on-device architectures; and the remainder contribute to modular, remixable ecosystems. This cluster frames “smallness” as a strategy for autonomy, sustainability, and openness—positioning small models as public infrastructure for high-performance AI outside the cloud. Focusing on accessibility and decentralised design philosophies, models and organisations in this cluster emphasise the compact nature of small models, which can be fine-tuned according to specific tasks and hosted offline and/or on-device. Energy-efficiency, low-resources and locality become central principles of this cluster, fostering modular ecosystems and enabling a community-driven registry to surface fine-tunes, prompt strategies, and inference hacks from diverse contributors.

transparent, legally consistent pipelines. These projects show that low-compute models can match, or even surpass, far larger proprietary systems while slashing costs. Embodying principles of responsible AI, safety, open-justice and democratizing access for education, civic tech and under-resourced regions - all without compromising transparency or user trust.



Modularity class four is titled *Green AI: Low-Footprint Models for Biodiversity*. This cluster reimagines small as field-ready, volunteer-run systems that privilege ethical AI, open science, and systems thinking. Spanning grassroots collectives, academic labs, and ranger stations, the projects translate AI research into boots-on-the-ground impact. The various projects allow rangers to monitor poaching hotspots (ICARUS), biologists to decode animal signals (NatureLM-Audio, Project CETI), and grassroots teams to democratise conservation knowledge without the carbon burden of giant models (Masakhane; BLOOM). Roughly one-third of the initiatives foreground real-time wildlife monitoring and threat detection, another third focus on open-access language or audio models that democratize conservation knowledge, and the remainder pioneer ethics-of-listening frameworks that treat non-human actors as stakeholders. From whale-song translators and telemetry-driven early-warning networks to sustainability analytics pipelines (ARIES) and fully open-weight language models, each effort demonstrates that lightweight architectures can power real-time biodiversity monitoring, facilitate inter-species understanding, and distribute ecological insights equitably. Rooted in interdisciplinary science, data governance, and respect for non-human actors, these slower, low-power tools listen first, act responsibly, and help secure a shared ecological future.



Modularity class five is titled *Indigenous-led Models for Sovereignty, Justice, and Cultural Continuity*. In this cluster, smallness itself is a political stance: AI as a locally governed, privacy-respecting companion that revitalises endangered languages, safeguards knowledge traditions, and advances justice on Indigenous terms. Grounded in principles of data sovereignty, kaitiakitanga, and reparative justice, the projects redefine small by intentionally scoping models to a single language, region, or community. The project landscape is broad yet coherent. Roughly half of the projects focus squarely on language revitalisation and cultural continuance (Divvun/Giellatekno; Kōrero Māori; Papa Reo) demonstrating how tightly scoped models can out-perform general systems on speech-to-text, text-to-speech, and corpus search for under-resourced tongues. About a third interrogate decolonial knowledge infrastructures and licensing (AI Decolonial Manifesto; O'Kanata), articulate governance frameworks. The remaining projects advance justice applications (MyShakti; SOF+IA). By training these models and toolkits on community-owned servers that run offline, these multilingual toolkits become accessible, low-cost instruments for cultural preservation, anti-violence advocacy, and regional equity rather than extractive pipelines. Ethical licenses such as Kaitiakitanga ensure communities control collection, use, and sharing of their corpora, while participatory design - elders curating story datasets, youth testing speech-to-text apps - keeps development aligned with Indigenous worldviews.

Pipeline step & key knobs

1. Package install

- bertopic, sentence-transformers, etc.

2. Data loading & cleaning

- JSON loading, filtering
- title + abstract concat
- deduplication

3. Embedding model

- Snowflake/snowflake-arctic-embed-s

4. UMAP

- n_neighbors, min_dist, n_components, metric, random_state

5. HDBSCAN

- min_cluster_size, metric, cluster_selection_method

6. CountVectorizer

- stop_words, ngram_range, max_df, min_df

7. BERTopic training

- top_n_words, calculate_probabilities

8. Outlier reduction

- strategy="probabilities"

As a final step, we provided the ChatGPT-o3 with the Vibe Mapping Co-lab code, and asked to provide a cheat sheet for fine-tuning the model on the basis of the various value-clusters. This resulted in a quick-reference grid that shows which parts of the BERTopic pipeline (steps 1 → 8) deserve special attention when you want to uphold all five value clusters we identified earlier. Each column lists the *parameters or settings you can tune and fiddle with*, tied to the values or cautions that flow from the Slow AI approach.

Open Source, Low-Compute & High-Performance

1. Avoid extra dependencies, choose light backends.
2. Drop duplicates and nulls early to shrink RAM load.
3. Try distilled or quantized models (MiniLM, int8 versions).
4. Lower n_components (e.g. 2–5) and increase min_dist to save compute.
5. Use cosine metric to avoid unnecessary distance matrix memory.
6. Set min_df=5, max_df=25 to control vocab size.
7. Turn off calculate_probabilities to reduce RAM.
8. Skip reduction for small datasets to save compute.

Small Data, Big Impact

1. Pin versions and pip freeze to support reproducibility; push to Git repo or Hugging Face Space.
2. Save original dataset hash and basic metadata in README.
3. Record model version, SHA, and license in logs or output.
4. Record every parameter setting for reuse and reproducibility.
5. Save cluster structure and plots for comparison.
6. Save final vocab as JSON; include in repo.
7. Set verbose=True; export logs and metrics to share.
8. Save pre-/post-outlier topic tables, document decisions.

Green AI: Low-Footprint Models for Biodiversity

1. Add energy-tracking libs (e.g. codecarbon, pyvmml).
2. Filter to domain-relevant records; tag with ecosystem or species code.
3. Replace with OpenAI3 or EcoAudio model for environmental audio embeddings.
4. Set 2D output and colour by timestamp for rhythm analysis (e.g. dawn chorus).
5. Trawl rare clusters as potential environmental signals; not noise.
6. Use character n-grams (3-5) if input is acoustic/symbolic or log-based.
7. Lower top_n_words to ~5 when working with ecological call data.
8. Outliers may be ecologically important anomalies—review carefully.

Indigenous-led Models for Sovereignty, Justice, and Cultural Continuity

1. Use .whl files and local installs if offline; avoid telemetry plugins.
2. Load only locally stored and community-consented files; avoid scraping.
3. Consider fine-tuning your own embedding model using culturally appropriate corpora.
4. Adjust metric and min_dist based on input from language/culture experts.
5. Iterate parameters with community stakeholders; document cluster logic.
6. Build your own culturally aware stop-word list; document choices.
7. Create merge/rename callbacks that invite participatory labeling.
8. Tag but retain rare, elder-contributed or unique records.

Feminist and Queer Approaches

1. Lock package versions so collaborators reproduce outputs; document setup decisions.
2. Check and flag gendered terms, bias, or underrepresented geographies; annotate rather than erase.
3. Choose bias-audited or inclusive models (e.g. all-mpnet-base-v2); check pronoun handling.
4. Use lower n_neighbors (4–6) to protect niche topic formation; set random_state=42.
5. Use smaller min_cluster_size (e.g. 3–6) to let marginal topics emerge.
6. Keep inclusive pronouns; avoid default stop-lists that erase gender, race, or class markers.
7. Use top_n_words=10 to surface nuanced language; retain probabilities to trace topic strength.
8. Manually review outliers before removal, especially for memorialized voices.

Methodologically, the translation of ethical values into actionable technical parameters represents a key innovation of this project. The structured approach of adjusting the BERTopic pipeline according to the values identified within each thematic cluster demonstrated that value-based fine-tuning is both feasible and meaningful. For instance, the Open Source, Low-Compute and High-Performance cluster operationalizes smallness through minimalistic computing choices—using distilled or quantized models, reducing dataset size early, and skipping computationally intensive steps like probability calculations

and outlier reduction. This approach concretely demonstrates how values of openness and computational sustainability can guide resource-conscious model deployment.

Similarly, the Feminist and Queer cluster operationalizes intersectional ethics by recommending parameter choices that explicitly address marginalization risks, such as lowering `min_cluster_size` to protect niche topics, manually reviewing outliers to safeguard marginalized voices, and carefully auditing for gendered biases through inclusive embeddings and annotating biased terms. Such parameter adjustments not only reflect theoretical commitments to inclusivity and representation but also embed these values directly into technical workflows.

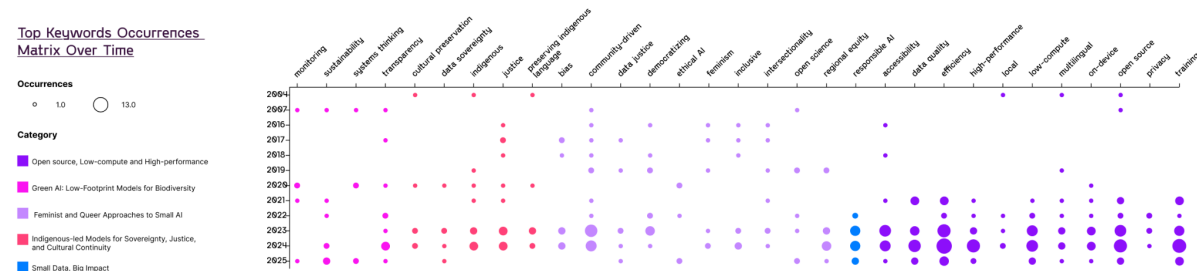
For Indigenous-led models, smallness operationalizes data sovereignty and cultural specificity through meticulous local data handling, offline-compatible formats (`.whl`), culturally appropriate embedding fine-tuning, and expert-driven parameter adjustments (e.g., altering `min_dist` according to linguistic and cultural expertise). This method exemplifies how cultural and political commitments can inform model choices to ensure community governance and ethical responsiveness.

The Green AI cluster specifically operationalizes sustainability through parameters such as energy tracking, ecological embedding models, domain-specific filters, and sensitivity to ecological anomalies. This precision in fine-tuning showcases how ecological responsibility can be embedded within AI workflows, aligning technical practice with environmental values.

Lastly, the Small Data, Big Impact cluster operationalizes transparency and accountability by standardizing reproducible environments, documenting dataset provenance, and ensuring detailed logging and metrics sharing. This rigorous transparency not only maintains user trust but also proves that ethical and responsible AI can be operationalized even within minimal data infrastructures.

Overall, these parameter guidelines deepen the argument that small AI principles are not merely aspirational—they can and should actively shape AI design decisions, transforming ethical, ecological, feminist, and Indigenous values into concrete, technically precise practices.

<insert poster section here>



6. Discussion

This study demonstrates that the concept of "small AI" encompasses a range of ethical, political, and technical values that extend beyond mere model size or computational efficiency. By identifying five thematic clusters, the analysis broadens the understanding of smallness within AI discourses, emphasizing its potential as a deliberate value-driven stance.

The Feminist and Queer cluster emphasizes intersectional ethics, care, and community-driven governance, actively challenging structural biases in mainstream AI. These initiatives adopt co-creative design practices, inclusive datasets, and participatory data governance methods, translating feminist and intersectional values into actionable technical measures like bias mitigation. Similarly, **the Small Data, Big Impact cluster** reframes "smallness" by prioritizing data quality, transparency, consent-driven processes, and civic or educational applications. This cluster illustrates how low-compute models can achieve significant societal impacts without sacrificing accuracy or user trust, challenging conventional assumptions linking model size to effectiveness. **The Open Source and Low-Compute Performance cluster** advocates decentralized, transparent, and energy-efficient AI infrastructures as alternatives to corporate-controlled models. Projects such as TinyLlama and llama.cpp demonstrate that lightweight, locally hosted, community-driven ecosystems can achieve high performance while democratizing access to AI tools. Meanwhile, the **Green AI** cluster integrates ecological ethics, systems thinking, and sustainability, promoting low-footprint AI solutions for biodiversity and environmental monitoring. Lastly, **Indigenous-led Models for Sovereignty** position small AI as politically and culturally significant, emphasizing data sovereignty, ethical licensing, and community governance as core elements that counteract colonial legacies, enabling AI to actively support cultural continuity and autonomy.

- some methodological reflections here, both on collaborating with GPT for data collection and curation (what does it add? when does it fail? how did we wrangle between manual and semi-automatic coding?)
- Also methodological reflections on what does our method add to current 'small' practices, when we propose translating principles from clusters into practical suggestions on how to tweak the model

7. Conclusion

8. References

A Comprehensive Survey of Small Language Models in the Era of Large Language Models: Techniques, Enhancements, Applications, Collaboration with LLMs, and Trustworthiness.

(n.d.). Retrieved July 4, 2025, from <https://arxiv.org/html/2411.03350v1#S1>

AfriBERTa brings the power of natural language processing to African languages | Cheriton

School of Computer Science | University of Waterloo. (n.d.). Retrieved July 4, 2025, from

<https://uwaterloo.ca/computer-science/news/afriberta>

Alzoubi, Y. I., & Mishra, A. (2024). Green artificial intelligence initiatives: Potentials and challenges. *Journal of Cleaner Production*, 468, 143090.

<https://doi.org/10.1016/j.jclepro.2024.143090>

Animals' early warning system. (n.d.). Retrieved July 4, 2025, from

<https://www.icarus.mpg.de/28810/animals-warning-sensors>

ARIES - ARtificial Intelligence for Environment & Sustainability. (n.d.). ARIES. Retrieved July 4, 2025, from <https://aries.integratedmodelling.org/>

Aymur AI: Measuring gender based violence in Latin America – A+ Alliance. (n.d.). Retrieved July 4, 2025, from

<https://aplusalliance.org/aymur-ai-measuring-gender-based-violence-in-latin-america/>

Beaujour, J.-M. (2024, November 30). PocketPal AI — Bringing Small Language Models to Your Cellphone. *Medium*.

<https://jmlbeaujour.medium.com/pocketpal-ai-bringing-small-language-models-to-your-cellphone-5ae486902f1b>

Beyond (the sea and) uncanny valleys and American smiles. (2024, March 25). *Alexia Achilleos*.

<https://alexiaachilleos.com/beyond-the-sea-and-uncanny-valleys-and-american-smiles/>

Black, E., Williams, J., Madaio, M. A., & Donti, P. L. (2020, April). A call for universities to develop requirements for community engagement in ai research. In *The Fair and Responsible AI Workshop at the 2020 CHI Conference on Human Factors in Computing Systems*.

Botchen, ein sehr kleiner Chatbot. (n.d.). Botchen, Ein Sehr Kleiner Chatbot. Retrieved July 4, 2025, from <https://botchen.org/>

Budenny, S., Lazarev, V., Zakharenko, N., Korovin, A., Plosskaya, O., Dimitrov, D., Akhripkin, V., Pavlov, I., Oseledets, I., Barsola, I., Egorov, I., Kosterina, A., & Zhukov, L. (2023). eco2AI: Carbon Emissions Tracking of Machine Learning Models as the First Step Towards Sustainable AI. *Doklady Mathematics*, 106. <https://doi.org/10.1134/S1064562422060230>

Bussler, F. (2024, February 13). The Surprising Rise of “Tiny AI.” *Predict*.

<https://medium.com/predict/the-surprising-rise-of-tiny-ai-59be99129e9e>

Can small language models revitalize Indigenous languages? (n.d.). *Brookings*. Retrieved July 4, 2025, from

<https://www.brookings.edu/articles/can-small-language-models-revitalize-indigenous-languages/>

Ciston, S. (2019). Intersectional AI is essential: Polyvocal, multimodal, experimental methods to save artificial intelligence. *Journal of Science and Technology of the Arts*, 11(2), Article 2.

<https://doi.org/10.7559/citarj.v11i2.665>

Corrêa, N. K., Falk, S., Fatimah, S., Sen, A., & De Oliveira, N. (2024). TeenyTinyLlama:

Open-source *tiny* language models trained in Brazilian Portuguese. *Machine Learning with Applications*, 16, 100558. <https://doi.org/10.1016/j.mlwa.2024.100558>

Costanza-Chock, S. (2018). Design Justice, A.I., and Escape from the Matrix of Domination.

Journal of Design and Science. <https://doi.org/10.21428/96c8d426>

Cyprus as AI Saw it: Reinventing the Colonial Gaze – Alexia Achilleos. (n.d.). Retrieved July 4,

2025, from <https://alexiaachilleos.com/cyprus-as-ai-saw-it-reinventing-the-colonial-gaze/>

Data Authenticity, Consent, and Provenance for AI Are All Broken: What Will It Take to Fix

Them? · From Novel Chemicals to Opera. (n.d.). Retrieved July 4, 2025, from

<https://mit-genai.pubpub.org/pub/uk7op8zs/release/2>

Decolonial AI as Disenclosure. (n.d.). Retrieved July 4, 2025, from

<https://www.scirp.org/journal/paperinformation?paperid=131479>

Decolonial AI Manyfesto. (n.d.). Retrieved July 4, 2025, from <https://manyfesto.ai/>

Denotation.eu. (n.d.). Retrieved July 4, 2025, from <https://www.denotation.eu/>

Fadelli, I. & Phys.org. (n.d.). *Shrinking AI for personal devices: An efficient small language model*

that could perform better on smartphones. Retrieved July 4, 2025, from

<https://techxplore.com/news/2024-11-ai-personal-devices-efficient-small.html>

Feminist AI. (n.d.). Feminist AI. Retrieved July 4, 2025, from <https://www.feminist.ai>

Feminist AI Papers, Prototypes & Pilots – A+ Alliance. (n.d.). Retrieved July 4, 2025, from

<https://aplusalliance.org/feminist-ai-papers/>

Fine-Tuning a Small Language Model for Secrets Detection | Wiz Blog. (2025, June 10). Wiz.io.

<https://www.wiz.io/blog/small-language-model-for-secrets-detection-in-code>

- GPT4All – The Leading Private AI Chatbot for Local Language Models*. (n.d.). Retrieved July 4, 2025, from <https://www.nomic.ai/gpt4all>
- Hassan, S. (2024, December 19). Google DeepMind Introduces “SALT”: A Machine Learning Approach to Efficiently Train High-Performing Large Language Models using SLMs. *MarkTechPost*.
<https://www.marktechpost.com/2024/12/19/google-deepmind-introduces-salt-a-machine-learning-approach-to-efficiently-train-high-performing-large-language-models-using-slms/>
- Jacaranda launches first-in-kind Swahili Large Language Model—Jacaranda Health*. (2023, October 31).
<https://jacarandahealth.org/jacaranda-launches-first-in-kind-swahili-large-language-model/>
- Jovanović, M., & Campbell, M. (2024). Compacting AI: In Search of the Small Language Model. *Computer*, 57(8), 96–100. <https://doi.org/10.1109/MC.2024.3404674>
- Klein, L., & D'Ignazio, C. (2024). Data Feminism for AI. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 100–112.
<https://doi.org/10.1145/3630106.3658543>
- Law, M. (2025, February 27). *Phi-4: Behind Microsoft's Smaller, Multimodal AI Models*.
<https://technologymagazine.com/articles/phi-4-behind-microsofts-smaller-multimodal-ai-models>
- Mashette, N. (2023, December 12). Small Language Models (SLMs). *Medium*.
<https://medium.com/@nageshmashette32/small-language-models-slms-305597c9edf2>
- Mensah, M., & van Wynsberghe, A. (2025). Sustainable AI meets feminist African ethics. *AI and Ethics*, 5(4), 4293–4303. <https://doi.org/10.1007/s43681-025-00705-9>
- Mohamed, S., Png, M.-T., & Isaac, W. (2020). Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology*, 33.
<https://doi.org/10.1007/s13347-020-00405-8>
- Nigeria, G. (2023, May 2). Firm to launch keyboard to communicate in 180 African indigenous languages. *The Guardian Nigeria News - Nigeria and World News*.
<https://guardian.ng/news/firm-to-launch-keyboard-to-communicate-in-180-african-indigenous-languages/>

- Pirie, C., Kalutarage, H., Hajar, M. S., Wiratunga, N., Charles, S., Madhushan, G. S., Buddhika, P., Wijesiriwardana, S., Dimantha, A., Hansamal, K., & Pathiranage, S. (2024). *A Survey of AI-Powered Mini-Grid Solutions for a Sustainable Future in Rural Communities* (arXiv:2407.15865). arXiv. <https://doi.org/10.48550/arXiv.2407.15865>
- Project CETI •—Home. (n.d.). Retrieved July 4, 2025, from <https://www.projectceti.org>
- Queer AI. (n.d.). Retrieved July 4, 2025, from <https://benlerchin.com/portfolio/queer-ai/>
- Queerinaï, O. O., Ovalle, A., Subramonian, A., Singh, A., Voelcker, C., Sutherland, D. J., Locatelli, D., Breznik, E., Klubicka, F., Yuan, H., J, H., Zhang, H., Shriram, J., Lehman, K., Soldaini, L., Sap, M., Deisenroth, M. P., Pacheco, M. L., Ryskina, M., ... Stark, L. (2023). Queer In AI: A Case Study in Community-Led Participatory AI. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1882–1895. <https://doi.org/10.1145/3593013.3594134>
- Quijano, Ricaurte, P., Zasso, M., & Polo, I. (2024). *Feminist AI in Latin America and the Caribbean: Building the future we want 2021-2024*. <https://hdl.handle.net/10625/63510>
- Razzaq, A. (2025, January 10). Meet KaLM-Embedding: A Series of Multilingual Embedding Models Built on Qwen2-0.5B and Released Under MIT. *MarkTechPost*. <https://www.marktechpost.com/2025/01/09/meet-kalm-embedding-a-series-of-multilingual-embedding-models-built-on-qwen2-0-5b-and-released-under-mit/>
- Richter, V., Katzenbach, C., & Schäfer, M. S. (2023). *Chapter 18: Imaginaries of artificial intelligence*. <https://www.elgaronline.com/edcollchap/book/9781803928562/book-part-9781803928562-24.xml>
- Sinders, C. (2020). Feminist data set. *Clinic for Open Source Arts*. <https://carolinesinders.com/wp-content/uploads/2020/05/Feminist-Data-Set-Final-Draft-2020-0517>
- Small Language Models: A Guide With Examples. (n.d.). Retrieved July 4, 2025, from <https://www.datacamp.com/blog/small-language-models>
- Small language models (SLMs): A cheaper, greener route into AI. (2023, November 30). <https://www.unesco.org/en/articles/small-language-models-slms-cheaper-greener-route-ai>
- Spawning. (n.d.). Spawning.Ai. Retrieved July 4, 2025, from <https://spawning.ai/>

Sun, Y., Tang, Y., Gao, Z., Pan, Z., Xu, C., Chen, Y., Qian, K., Wang, Z., Braud, T., Lee, C. H., & Asadipour, A. (2023). *AI Nushu: An Exploration of Language Emergence in Sisterhood -Through the Lens of Computational Linguistics* (arXiv:2310.11870). arXiv. <https://doi.org/10.48550/arXiv.2310.11870>

Tech 4 Tracing. (n.d.). Retrieved July 4, 2025, from <https://tech4tracing.org/>

The Noosphere manifested: AI as instrument of knowledge extractivism | AI & SOCIETY. (n.d.). Retrieved July 4, 2025, from <https://link.springer.com/article/10.1007/s00146-020-01097-6>

The Tiny LLM Revolution—Part 1. (2024, January 4). Prem. <https://blog.prem.ai/the-tiny-llm-revolution-part1/>

Top 15 Small Language Models for 2025. (n.d.). Retrieved July 4, 2025, from <https://www.datacamp.com/blog/top-small-language-models>

Trustworthy AI: AI made in Germany and Europe? | AI & SOCIETY. (n.d.). Retrieved July 4, 2025, from <https://link.springer.com/article/10.1007/s00146-023-01808-9>

Tu, X., Chen, D., Han, K., Altintas, O., & Wang, H. (2025). *GreenAuto: An Automated Platform for Sustainable AI Model Design on Edge Devices* (arXiv:2501.14995). arXiv. <https://doi.org/10.48550/arXiv.2501.14995>

Varon (Joana), & Peña (Paz). (2021, December 7). *Artificial intelligence and consent: A feminist anti-colonial critique* [Info:eu-repo/semantics/article]. Alexander von Humboldt Institute for Internet and Society gGmbH. <https://doi.org/10.14763/2021.4.1602>

Varshney, K. R. (2024). *Decolonial AI Alignment: Openness, Viśe\{s\}a-Dharma, and Including Excluded Knowledges* (arXiv:2309.05030). arXiv. <https://doi.org/10.48550/arXiv.2309.05030>

VFRAME. (n.d.). Retrieved July 4, 2025, from <https://vframe.io/>

Welcome to Chechnya. (n.d.). Retrieved July 4, 2025, from <https://www.welcometochechnya.com/>

What are Small Language Models (SLM)? | IBM. (n.d.). Retrieved July 4, 2025, from <https://www.ibm.com/think/topics/small-language-models>

With genAI models, size matters (and smaller may be better) – Computerworld. (n.d.). Retrieved July 4, 2025, from <https://www.computerworld.com/article/3529501/smaller-genai-models-for-every-app-might-be-the-future.html>

Yotta Launches myShakti Gen AI Chatbot that Runs Off DeepSeek Open-Source AI Model.

(n.d.). Retrieved July 4, 2025, from

<https://techachievemedia.com/press-release/yotta-launches-myshakti-gen-ai-chatbot-deepseek-open-source-ai-model/>

Yuqian Sun | CVPR AI Art. (n.d.). Retrieved July 4, 2025, from

<https://thecvf-art.com/project/ai-nushu-ai%E5%A5%B3%E4%B9%A6-2023/>

9. Related projects