Extended: Using Large Language Models to enhance granularity of biographical data

Philipp Koch
EcoAustria – Institute for Economic Research, Vienna, Austria

Abstract

Biographical records of famous individuals offer a comprehensive window into the economic and cultural geography of the past. Yet, current datasets built from these records often lack granularity. Oversimplified migration histories and occupational classifications limit their potential for economic and historical research. Here, we propose leveraging state-of-the-art Large Language Models (LLMs) to systematically extract detailed information from biographies available on Wikipedia and Wikidata that other sources did not yet retrieve. Specifically, we aim to capture precise migration trajectories, multiple occupations throughout individuals' careers, interpersonal connections, and university affiliations. By integrating these aspects into Pantheon (pantheon.world), a platform run by César A. Hidalgo and his colleagues at Datawheel that attracts more than 2 million users per year, we will enable researchers and the public to explore historical knowledge diffusion, occupational mobility, and regional economic development. Also, we use the retrieved data to augment historical population data for regions through machine learning methods. Together, the granular biographical data we retrieve and make publicly available will provide novel insights into the determinants of economic prosperity and long-term growth.

Introduction

Biographies of notable individuals are a highly comprehensive representation of the historical economic geography. Our collective memory on the careers of Michelangelo, Wolfgang Amadeus Mozart, Alexander Fleming, or Albert Einstein sheds light on the economic and cultural ecosystem of their respective locations in ways traditional archival sources cannot. Data on such biographies is, hence, an increasingly used source in economic, historical, and interdisciplinary research (1–6).

Unlike typical historical data sources, there is an abundance of accurate biographical records—largely thanks to community-driven efforts such as Wikipedia and Wikidata, which provide extensive biographical information at an unprecedented scale. Recent research efforts used this as a starting point and made available structured data on the places of birth, death, and occupations of hundreds of thousands of historical figures (7, 8).

But these research efforts lack crucial information on the biographies of famous individuals, limiting the research questions that can be explored.

Take migration patterns. Most of the migration of famous Europeans over the past 1,000 years took place within countries and towards large cities (e.g. from smaller cities in France to Paris, see Fig. 1). And they were remarkably mobile

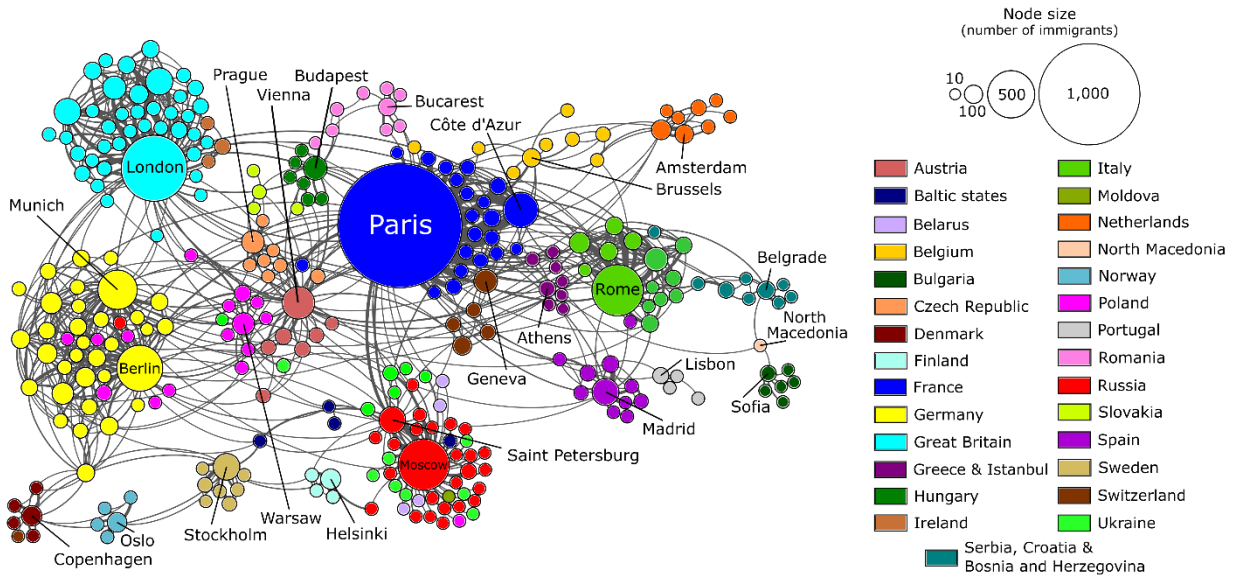

Figure 1. *Migration network of famous individuals within Europe between the years 1000 and 2000. Places of birth and death are used as a proxy for migration. Capturing all migration flows would increase the number of links in this graph substantially.*

throughout history. Scientists born between 1450 and 1750 moved, on average, 3.7 times in their lives (9). But curated datasets (7, 8) only provide information on their places of birth and death, which researchers then use as a proxy of migration (2–4). Although this proxy is appropriate, it leaves out relevant details. With this proxy, Albert Einstein is considered a migrant of Ulm to Princeton, neglecting his stays in Munich, Zürich, Bern, and Berlin. Information on all places of living and granular information on the individuals' migration trajectory, however, is contained in long-format biographies and should be mineable in a structured way.

Similarly, current datasets assign a single occupation to each famous individual, which can be oversimplified. Consider the career of Maria Sibylla Merian. She started her career as a botanical artist, illustrating flowers and insects. Later, she contributed significantly to the emerging field of entomology with several scientific observations. Capturing such labor flows systematically can inform us about the cognitive proximity between two occupations (e.g. botanical artist and entomologist as showcased by Merian's biography). This, in turn, can enhance our understanding of economic development and knowledge diffusion through economic complexity methods such as relatedness (4, 10, 11). Other properties, which are contained in long-format biographies but are not yet captured systematically, are connections between famous individuals (e.g. ontological inspiration, mentorship, friendship etc.) or their affiliation to universities.

Here, we want to address these limitations by implementing state-of-the-art Large Language Models (LLMs) with Wikidata and the long-format biographies on Wikipedia. LLMs are powerful classification tools which find structure and discrete classes within vast amounts of unstructured data (12, 13). This enables completely new use cases of text as data.

The goal of the project is to retrieve more granular information on the famous individuals' places of living and migration patterns, their occupational changes over their lifetime, the connections between famous individuals, and their affiliation to universities.

The expected output of this project are a new dataset including above mentioned aspects, two academic publications, and the integration of the retrieved data into Pantheon.

The first publication will summarize our methodological approach and make the collected data on famous individuals publicly available. We will ensure compatibility with existing datasets and plan to collaborate with the Wikimedia Foundation to make our findings directly accessible on Wikidata.

Also, we will closely collaborate with César A. Hidalgo and his colleagues at Datawheel to integrate our approach within the pipeline of pantheon.world and generate a new version of Pantheon (Pantheon 3.0). Pantheon is a popular platform, attracting more than 2 million users per year from all around the world. This makes our data accessible to researchers and the general public.

The second publication targets the augmentation of historical population data through machine learning, similar to our recent work on augmenting the availability of historical GDP per capita levels (5).

Date: We propose a two-year project starting in July 2025 and ending in June 2027.

Related work

Using biographies of famous individuals to understand cultural and economic development is not an arbitrary choice. Accurate information on their lives is abundant compared to traditional historical data, and it contains relevant details to understand the development and structural change of cities.

Key agents of structural change are migrants (14–18). They help carry knowledge across space and shape the geography of cultural and economic activities (19–23). Several studies documented the historical role of migrants in knowledge diffusion within a single activity (24–34). For instance, the forced emigration of French Hugenots, who settled in Prussia in the late 17th century, significantly increased textile manufacturing productivity in Prussia (35).

The role of migrants in the historical formation of agglomerations spanning multiple activities, however, remained relatively unexplored. In a recent publication (4), we explore their role with data on more than 22,000 famous individuals living in Europe in the past 1,000 years. Specifically, we investigated how immigrants, emigrants, and locals explain the probability that famous individuals specialized in an activity—that was not yet present in a region—are born during the next century. Put differently: Did migrants make Paris a Mecca for the arts and Vienna a beacon of classical music? Or was their rise a pure consequence of local actors?

Our findings show that migrants play a crucial role in the historical geography of knowledge. Specifically, we find that the probability that a European region enters a new activity grows with the presence of immigrants with knowledge on that activity. Also, using measures of relatedness (11, 36–39), we find that this correlation is enhanced by spillovers across related activities. Put differently, the probability that a region begets famous mathematicians grows with an excess immigration of mathematicians and with immigrants from related fields, such as physics or chemistry. Similarly, we find that the probability that a European region loses one of its existing areas of specialization decreases with the presence of immigrants specialized in that activity and in related activities. In contrast, we do not find that locals play a statistically significant role in shaping European cities.

But biographies of famous individuals can also be used to augment the availability of historical economic data such as GDP per capita levels with the help of machine learning. In a recent publication (5), we introduce a machine learning method designed to reconstruct historical GDP

per capita estimates of dozens of European and North American countries and regions for the past 700 years, more than quadrupling the availability of historical economic output data for these regions.

Leveraging information on 563,000 famous historical figures from Wikipedia and Wikidata (8), we trained a set of supervised machine learning models to generate out-of-sample estimates of national and regional GDPs per capita. We find the model provides encouraging results. In an out-of-sample test, it predicts the GDP per capita of European and North American countries and regions with an R^2=90.1% and a mean absolute error of 22.6% of the GDP per capita observed during that time period.

We externally validate these estimates by recreating qualitatively well-known historical development trajectories and by comparing them with other proxies of per capita wealth. First, we recreate the established finding that England and the Low Countries experienced larger economic growth than Southern Europe between 1300 and 1800 (40–44). We find that a large share of this reversal of fortune can be attributed to the rise of Atlantic trade (40). Second, we show our estimates correlate with proxies of economic development, such as urbanization rates between 1500 and 1950, body height in the 18th century, wellbeing in 1850, and church building activity in the 14th and 15th century.

Our articles on migrants as drivers of historical structural change and on augmenting the availability of historical GDP per capita data enhance our understanding of economic development.

But the underlying data on famous historical figures suffers from key shortcomings we want to tackle with this project.

First, we lack granular information on the full migration trajectory of famous individuals. Rather, places of birth and death are used as a proxy for migration. More detailed data on where famous individuals lived, and when, could provide a better analytical basis to explore the evolution of agglomerations. This is relevant for both our papers cited above. We could provide better estimates of the impact of migrants on structural change, if we observed all places of living. Similarly, we could improve the model performance of machine learning augmenting historical economic data such as GDP per capita levels.

Second, the assignment of occupations to biographies can be improved. Current research efforts assign a single occupation to an individual, while many famous figures contributed to several fields in the course of their career. Consider Maria Sibylla Merian, who started her career as botanical artist and ended it as entomologist. Or Leonardo da Vinci, whose Wikipedia page describes him as "painter, draughtsman, engineer, scientist, theorist, sculptor, and architect."

A more granular and more accurate assignment of occupations can enhance our understanding of knowledge diffusion. When exploring the role of migrants in the evolution of European agglomerations, we used measures of relatedness (11, 36–39) to capture knowledge spillovers across activities. Relatedness estimates how cognitively close a location is to an activity. The starting point in our case was a network of proximities between activities based on their co-location. That is, we assume that, e.g., painters and sculptors are cognitively close when they frequently appear in the same location. Having granular information on occupations a single individual pursues can provide a more accurate depiction of how cognitively close two occupations are. This, in turn, can lead to improved measures of relatedness and knowledge spillovers.

Together, we propose addressing these limitations by implementing state-of-the-art Large Language Models (LLMs) with Wikidata and the long-format biographies on Wikipedia.

Interdisciplinary research using data on the biographies of famous individuals can profit from these advancements and enhance our understanding of past economic development even further.

Methods

This project starts with the current pipeline that is feeding the website pantheon.world. Pantheon started as a project by MIT's Collective Learning group, which resulted in Pantheon 1.0 using data from Wikipedia and Wikidata (7). Then, Datawheel continued the development of Pantheon and expanded its scope to today's version, Pantheon 2.0. Now, more than 2 million users visit Pantheon per year.

We will work closely with César A. Hidalgo and colleagues at Datawheel to benefit from knowledge spillovers and implement LLMs into their pipeline, eventually contributing to a new version of Pantheon (Pantheon 3.0).

At the core of the project is the integration of LLMs into the workflow behind Pantheon. LLMs are powerful classification tools which find structure and discrete classes within vast amounts of unstructured data (12, 13). Which tasks LLMs can reliably solve and how such models can be adapted for different use cases and data types, is currently an active and growing field of research in computer science (45–53).

The first step of the project is a thorough literature review of current applications of LLMs for similar information retrieval use cases. The applied model (or ensemble of multiple models) will depend on the insights gained during the literature review. We will also take earlier attempts to retrieve more granular information from Wikidata into account, which used early language processing techniques and had limits in scale (54, 55).

After the literature review, we set up LLMs for the purpose of retrieving granular information from Wikipedia by fine-tuning relevant parameters and developing sets of prompts that reliably yield the respective item in a structured form.

The goal is to retrieve more granular information on:

- the famous individuals' places of living and migration patterns,
- their (potentially multiple and changing) occupations over their lifetime,
- the connections between famous individuals, and
- their affiliation to universities.

Next, we will use the retrieved data to augment historical population data. Historical population data is more prevalent than income data, but coverage in some parts of the world is still scarce. Here, we augment the availability of historical population data through machine learning with a similar approach to our work on GDP per capita levels (5). We will employ sets of supervised machine learning models (e.g. elastic net regression models or regression trees) and train them with several features derived from the geography of famous biographies.

Expected output

The expected output of the project includes:

- A publication in an interdisciplinary peer-reviewed journal describing our methodological approach in detail. This includes a publicly accessible dataset that interested researchers can use. We will ensure direct compatibility with existing data sources (7, 8) and Wikidata;
- A publication in an interdisciplinary peer-reviewed journal augmenting historical population data;
- Presentations at two academic conferences;
- Integrating our data within the pipeline of pantheon.world and generating a new version of Pantheon (Pantheon 3.0) to

make our data accessible to researchers and the general public;

Risks

Risks in this project relate to biases that data from Wikipedia is known to be subject to (56). For instance, famous figures of the Western world are overrepresented (57). Also, cultural norms impact the portrayal of certain individuals in different language editions (58, 59), and the relative coverage of topics (60). Still, empirical studies find that the information available in Wikipedia is of relatively high accuracy when assessed by experts (61) or compared with other encyclopedias such as Britannica (62). Additionally, findings need to be set in context, since they are based on a small, and not necessarily representative, subset of the overall population. All parts of this project will consider potential biases and limitations throughout, critically discuss them, and make sure that results are not driven by them.

Community impact plan

We will make our results accessible to the general public by closely collaborating with colleagues at Datawheel and publishing our retrieved information within Pantheon (pantheon.world).

Pantheon is a popular platform, attracting more than 2 million users per year from all around the world. While a large share of users (~30 percent) are based in the United States, the site is also popular in non-Western countries such as India (~6,5 percent of Pantheon users), China (~2,8 percent), and Nigeria (~1,9 percent).

Also, we will ensure compatibility with existing datasets and plan to collaborate with the Wikimedia Foundation to make our findings directly accessible on Wikidata.

Evaluation

This project will be evaluated based on several clearly defined criteria that align with its primary objectives.

First, success will be measured by the accuracy, completeness, and granularity of the information extracted using LLMs. Specifically, we will validate the retrieved data on famous individuals' residential histories, occupational trajectories, connections between individuals, and university affiliations against manually curated benchmark datasets. Also, we will conduct random manual checks on a representative subset of retrieved biographies to assess accuracy rates.

Second, a key indicator of success will be the integration of the new structured data into existing platforms, particularly Pantheon 3.0. We will track adoption and usage statistics of our datasets through access statistics and citations.

Third, publication of the project's methodology and results in reputable interdisciplinary peer-reviewed journals will serve as a critical measure of success. Depending on the speed of progress and the review times, a measure of success can already be the reception of positive referee reports with the possibility to revise the submitted manuscript.

Collectively, these evaluation criteria will ensure that the project's outcomes are robust, methodologically sound, and impactful within both academic research communities and public dissemination channels.

Budget

The total budget for this project amounts to USD 149 444,79. It is available here.

Personnel costs amount to USD 92 102,63. These include 12 full-time months of the applicant, as well as six full-time months of a research assistant (Master's student), who will be hired at EcoAustria in Vienna after approval of the project. These costs are calculated based on the

current standard personnel costs by the Austrian Science Fund FWF (FWF webpage).

Costs were converted from EUR to USD using the OANDA Currency Converter on April 14th resulting in an exchange rate of 1,13549 USD/EUR.

The project will be led and supervised by Philipp Koch at EcoAustria – Institute for Economic Research in Vienna. EcoAustria is an independent non-profit research institute in Vienna, covering academic research as well as current topics of public interest. The team members are highly experienced specialists with longstanding expertise in applied and academic economic research.

The project's PI, Philipp Koch, completed his PhD in 2024 at the Center for Collective Learning at the University of Toulouse and currently holds a position as Head of Data Science at EcoAustria. His dissertation of "Machine Learning for Economic History" resulted in two publications contributing to a better understanding of economic history (see section "Related work").

A central element of the project is to promote the career of young scientists. First, the PI only recently completed his doctoral degree. Also, we will hire a Master's student shortly after the project start, who supports the project in several ways. Specifically, we plan to attract students in Data Science to help with the computational aspect of the project.

A key aspect of this project is the public dissemination of the data and results to a large audience. Hence, we aim to closely collaborate with colleagues at Datawheel, who run the popular platform pantheon.world. Having Datawheel as subcontractor on board will help the proposed project have a high community impact. A budget of USD 45 419,53 is reserved for Datawheel to help facilitate the integration into their data pipeline and website. Also, software support by Datawheel will be provided if needed.

References

1. C. Jara-Figueroa, A. Z. Yu, C. A. Hidalgo, How the medium shapes the message: Printing and the rise of the arts and sciences. *PLOS ONE* **14**, e0205771 (2019).
2. M. Serafinelli, G. Tabellini, Creativity over time and space: A historical analysis of European cities. *J. Econ. Growth* **27**, 1–43 (2022).
3. M. Schich, *et al.*, A network framework of cultural history. *Science* **345**, 558–562 (2014).
4. P. Koch, V. Stojkoski, C. A. Hidalgo, The Role of Immigrants, Emigrants, and Locals in the Historical Formation of European Knowledge Agglomerations. *Reg. Stud.* **58**, 1659–1673 (2023).
5. P. Koch, V. Stojkoski, C. A. Hidalgo, Augmenting the availability of historical GDP per capita estimates through machine learning. *Proc. Natl. Acad. Sci.* **121**, e2402060121 (2024).
6. D. De La Croix, O. Licandro, The longevity of famous people from Hammurabi to Einstein. *J. Econ. Growth* **20**, 263–303 (2015).
7. A. Z. Yu, S. Ronen, K. Hu, T. Lu, C. A. Hidalgo, Pantheon 1.0, a manually verified dataset of globally famous biographies. *Sci. Data* **3**, 150075 (2016).
8. M. Laouenan, *et al.*, A cross-verified database of notable people, 3500BC-2018AD. *Sci. Data* **9**, 290 (2022).
9. J. Mokyr, Mobility, Creativity, and Technological Development: David Hume, Immanuel Kant and the Economic Development of Europe. [Preprint] (2005). Available at: https://faculty.wcas.northwestern.edu/jmokyr/Berlin.PDF.
10. F. Neffke, M. Henning, Skill relatedness and firm diversification. *Strateg. Manag. J.* **34**, 297–316 (2013).
11. C. A. Hidalgo, Economic complexity theory and applications. *Nat. Rev. Phys.* **3**, 92–113 (2021).
12. E. Ash, S. Hansen, Text Algorithms in Economics. *Annu. Rev. Econ.* **15**, 659–688 (2023).

13. M. Dell, Deep Learning for Economists. *J. Econ. Lit.* (2025). https://doi.org/10.48550/arXiv.2407.15339.
14. E. Miguelez, A. Morrison, Migrant Inventors as Agents of Technological Change. *J. Technol. Transf.* (2022). https://doi.org/10.1007/s10961-022-09927-z.
15. Z. Elekes, R. Boschma, B. Lengyel, Foreign-owned firms as agents of structural change in regions. *Reg. Stud.* **53**, 1603–1613 (2019).
16. F. Neffke, M. Hartog, R. Boschma, M. Henning, Agents of Structural Change: The Role of Firms and Entrepreneurs in Regional Diversification. *Econ. Geogr.* **94**, 23–48 (2018).
17. L. Putterman, D. N. Weil, Post-1500 Population Flows and The Long-Run Determinants of Economic Growth and Inequality. *Q. J. Econ.* **125**, 1627–1682 (2010).
18. A. Morrison, Towards an evolutionary economic geography research agenda to study migration and innovation. *Camb. J. Reg. Econ. Soc.* rsad013 (2023). https://doi.org/10.1093/cjres/rsad013.
19. F. Lissoni, International migration and innovation diffusion: an eclectic survey. *Reg. Stud.* **52**, 702–714 (2018).
20. M. Trippl, G. Maier, "Knowledge Spillover Agents and Regional Development" in *Innovation, Growth and Competitiveness*, Advances in Spatial Science., P. Nijkamp, I. Siedschlag, Eds. (Springer Berlin Heidelberg, 2011), pp. 91–111.
21. A. M. Williams, Lost in translation? International migration, learning and knowledge. *Prog. Hum. Geogr.* **30**, 588–607 (2006).
22. S. P. Kerr, W. Kerr, Ç. Özden, C. Parsons, High-Skilled Migration and Agglomeration. *Annu. Rev. Econ.* **9**, 201–234 (2017).
23. C. M. Cipolla, The Diffusion of Innovations in Early Modern Europe. *Comp. Stud. Soc. Hist.* **14**, 46–52 (1972).
24. P. Moser, A. Voena, F. Waldinger, German Jewish Émigrés and US Invention. *Am. Econ. Rev.* **104**, 3222–3255 (2014).
25. I. Ganguli, Immigration and Ideas: What Did Russian Scientists "Bring" to the United States? *J. Labor Econ.* **33**, S257–S288 (2015).
26. D. Diodato, A. Morrison, S. Petralia, Migration and invention in the Age of Mass Migration. *J. Econ. Geogr.* **22**, 477–498 (2022).
27. K. J. Borowiecki, K. Graddy, Immigrant artists: Enrichment or displacement? *J. Econ. Behav. Organ.* **191**, 785–797 (2021).
28. S. Mitchell, London calling? Agglomeration economies in literature since 1700. *J. Urban Econ.* **112**, 16–32 (2019).
29. K. J. Borowiecki, Are composers different? Historical evidence on conflict-induced migration (1816-1997). *Eur. Rev. Econ. Hist.* **16**, 270–291 (2012).
30. F. Waldinger, Quality Matters: The Expulsion of Professors and the Consequences for PhD Student Outcomes in Nazi Germany. *J. Polit. Econ.* **118**, 787–831 (2010).
31. F. Waldinger, Peer Effects in Science: Evidence from the Dismissal of Scientists in Nazi Germany. *Rev. Econ. Stud.* **79**, 838–861 (2012).
32. W. C. Scoville, The Huguenots and the Diffusion of Technology. I. *J. Polit. Econ.* **60**, 294–311 (1952).
33. W. C. Scoville, The Huguenots and the Diffusion of Technology. II. *J. Polit. Econ.* **60**, 392–411 (1952).
34. H. M. Collins, The TEA Set: Tacit Knowledge and Scientific Networks. *Sci. Stud.* **4**, 165–185 (1974).
35. E. Hornung, Immigration and the Diffusion of Technology: The Huguenot Diaspora in Prussia. *Am. Econ. Rev.* **104**, 84–122 (2014).
36. C. A. Hidalgo, B. Klinger, A.-L. Barabási, R. Hausmann, The product space conditions the development of nations. *Science* **317**, 482–487 (2007).
37. C. A. Hidalgo, *et al.*, "The Principle of Relatedness" in *Unifying Themes in Complex Systems IX*, Springer Proceedings in Complexity., A. J. Morales, C. Gershenson, D. Braha, A. A. Minai, Y. Bar-Yam, Eds. (Springer International Publishing, 2018), pp. 451–457.
38. R. Boschma, Relatedness as driver of regional diversification: a research agenda. *Reg. Stud.* **51**, 351–364 (2017).

39. P.-A. Balland, *et al.*, The new paradigm of economic complexity. *Res. Policy* **51**, 104450 (2022).
40. D. Acemoglu, S. Johnson, J. Robinson, The Rise of Europe: Atlantic Trade, Institutional Change, and Economic Growth. *Am. Econ. Rev.* **95**, 546–579 (2005).
41. A. M. de Pleijt, J. L. van Zanden, Accounting for the "Little Divergence": What drove economic growth in pre-industrial Europe, 1300–1800? *Eur. Rev. Econ. Hist.* **20**, 387–409 (2016).
42. A. Henriques, N. Palma, Comparative European Institutions and the Little Divergence, 1385–1800. *J. Econ. Growth* (2022). https://doi.org/10.1007/s10887-022-09213-5.
43. M. Fochesato, Origins of Europe's north-south divide: Population changes, real wages and the 'little divergence' in early modern Europe. *Explor. Econ. Hist.* **70**, 91–131 (2018).
44. R. C. Allen, The Great Divergence in European Wages and Prices from the Middle Ages to the First World War. *Explor. Econ. Hist.* **38**, 411–447 (2001).
45. S. Kim, *et al.*, Evaluating Language Models as Synthetic Data Generators. [Preprint] (2024). Available at: http://arxiv.org/abs/2412.03679 [Accessed 27 December 2024].
46. Y. Wang, *et al.*, Self-Instruct: Aligning Language Models with Self-Generated Instructions in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Association for Computational Linguistics, 2023), pp. 13484–13508.
47. O. Honovich, T. Scialom, O. Levy, T. Schick, Unnatural Instructions: Tuning Language Models with (Almost) No Human Labor in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Association for Computational Linguistics, 2023), pp. 14409–14428.
48. R. Liu, *et al.*, Best Practices and Lessons Learned on Synthetic Data. [Preprint] (2024). Available at: http://arxiv.org/abs/2404.07503 [Accessed 8 January 2025].
49. V. Viswanathan, C. Zhao, A. Bertsch, T. Wu, G. Neubig, Prompt2Model: Generating Deployable Models from Natural Language Instructions in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, (Association for Computational Linguistics, 2023), pp. 413–421.
50. S. Kim, S. Joo, Y. Jang, H. Chae, J. Yeo, CoTEVer: Chain of Thought Prompting Annotation Toolkit for Explanation Verification. *Proc. 17th Conf. Eur. Chapter Assoc. Comput. Linguist. Syst. Demonstr.* 195–208 (2023).
51. C. Xu, *et al.*, WizardLM: Empowering Large Language Models to Follow Complex Instructions. [Preprint] (2023). Available at: http://arxiv.org/abs/2304.12244 [Accessed 8 January 2025].
52. M. Rybinski, W. Kusa, S. Karimi, A. Hanbury, Learning to match patients to clinical trials using large language models. *J. Biomed. Inform.* **159**, 104734 (2024).
53. M. Staudinger, W. Kusa, F. Piroi, A. Lipani, A. Hanbury, A Reproducibility and Generalizability Study of Large Language Models for Query Generation in *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, (ACM, 2024), pp. 186–196.
54. L. Lucchini, S. Tonelli, B. Lepri, Following the footsteps of giants: modeling the mobility of historically notable individuals using Wikipedia. *EPJ Data Sci.* **8**, 36 (2019).
55. S. Menini, *et al.*, RAMBLE ON: Tracing Movements of Popular Historical Figures in *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, (Association for Computational Linguistics, 2017), pp. 77–80.
56. C. Hube, Bias in Wikipedia in *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, (ACM Press, 2017), pp. 717–721.
57. M. Dittus, M. Graham, Mapping Wikipedia's Geolinguistic Contours. *Digit. Cult. Soc.* **5**, 147–164 (2019).

58. U. Pfeil, P. Zaphiris, C. S. Ang, Cultural Differences in Collaborative Authoring of Wikipedia. *J. Comput.-Mediat. Commun.* **12**, 88–113 (2006).
59. E. S. Callahan, S. C. Herring, Cultural bias in Wikipedia content on famous persons. *J. Am. Soc. Inf. Sci. Technol.* **62**, 1899–1915 (2011).
60. A. Halavais, D. Lackaff, An Analysis of Topical Coverage of Wikipedia. *J. Comput.-Mediat. Commun.* **13**, 429–440 (2008).
61. T. Chesney, An empirical examination of Wikipedia's credibility. *First Monday* **11** (2006).
62. J. Giles, Internet encyclopaedias go head to head. *Nature* **438**, 900–901 (2005).