From Keywords to Structured Summaries: Streamlining Scholarly Information Access

Mahsa Shamsabadi, Jennifer D'Souza

TIB Leibniz Information Centre for Science and Technology, Hannover, Germany

Abstract

This poster paper highlights the increasing importance of information retrieval (IR) engines in the scientific community, addressing the inefficiencies of traditional keyword-based search engines amid the growing volume of publications. Our proposed solution uses structured records, supported by advanced information technology (IT) tools such as visualization dashboards, to transform how researchers access and filter articles, moving away from a text-heavy approach. This vision is demonstrated through a proof of concept focused on the "reproductive number estimate of infectious diseases" research theme. We utilize a fine-tuned large language model (LLM) to automate the creation of structured records for a backend database, enhancing information access beyond simple keywords. The result is a next-generation information access system, available at https://orkg.org/usecases/r0-estimates.

Keywords

Structured scientific knowledge, Structured scientific information extraction (IE), Large Language Models, Visualization dashboards, Scientific information retrieval (IR) platforms

1. Introduction

The rapid expansion of scientific literature necessitates a reevaluation of their information retrieval (IR) engines [1, 2]. Traditional keyword-based approaches are inadequate for tracking fast-paced scientific advancements. There is a growing demand for structured scientific content representations [3, 4] and advanced machine learning algorithms [5, 6] to enhance retrieval accuracy. Initiatives like the Open Research Knowledge Graph (ORKG) [7] drive this paradigm shift towards structured knowledge representations, enabling intelligent views and comparisons of research facets [8, 9]. Our goal is to simplify access to scientific articles and reduce cognitive load for researchers using information technology (IT). We propose dashboards as visual tools to represent structured scientific knowledge, enhancing research filtering and discovery processes [10]. Dashboards have been widely used, including during the Covid-19 pandemic, where they helped track cases, analyze trends, and support decision-making with data from sources like the WHO and Johns Hopkins [11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36]. In contrast, our approach focuses on applying IT to structure scientific knowledge itself, using information extraction (IE) mechanisms and large language models (LLMs) to power next-generation information systems.

Posters, Demos, and Industry Tracks at ISWC 2024, November 13–15, 2024, Baltimore, USA

†This work was supported by the German BMBF project SCINEXT (ID 01lS22070).

jennifer.dsouza@tib.eu (J. D'Souza)

0000-0002-6616-9509 (J. D'Souza)

In pursuit of our vision, this poster paper presents a proof of concept (POC) using the ORKG-R0 semantic model [37] to structure articles on the ”reproductive number estimate of infectious diseases” theme [38]. The model captures essential properties like disease name, study location, date, $R0$ value, % confidence interval values, and computation method, enabling effective comparison across studies. Four research questions (RQs) guide article search and exploration: **RQ1** identifies maximum $R0$ estimates, **RQ2** examines study counts by disease and location, **RQ3** analyzes $R0$ value ranges by location for selected diseases, and **RQ4** maps study locations globally on the world map. These RQs are visualized in a dashboard to enhance article filtering and provide researchers with concise insights into research progress.

2. Next-Generation Scientific Information Retrieval (IR)

We introduce a next-generation IR platform for "reproductive number estimates of infectious diseases," enhancing scientific article access with IT and four visual charted summaries tailored to four specific RQs alluded to earlier. In the following subsections, we will detail the LLM-based IE method, article collection, and platform workflow.

2.1. The Scientific Information Extraction (IE) Large Language Model (LLM)

We employ the ORKG-FLAN-T5 R0 LLM [39]. This model is an instruction fine-tuned variant of FLAN-T5 Large (780 M) using the instruction-tuning paradigm introduced as FLAN (Finetuned Language Net) [40, 41, 42, 43]. It processes a paper's title and abstract to produce structured summaries based on six key properties: *disease name*, *location*, *date*, *$R0$ value*, *% confidence interval (CI) values*, and *method*, related to the $R0$ estimate [39].

Table 1
The top 20 infectious disease names (and number of papers) in our initial dataset.

covid-19 (1002)	mers-cov (21)	measles (15)	hepatitis c (8)
dengue (41)	cholera (18)	hepatitis b (12)	tuberculosis (8)
influenza (29)	zika (18)	zika virus (12)	monkeypox (8)
hiv (23)	african swine fever (17)	ebola (11)	west nile virus (7)
sars (22)	ebola (17)	hand, foot, and mouth disease (8)	malaria (7)

2.2. The Scholarly Articles Collection

The initial set of articles in our collection was sourced from keyword-based searches in the PubMed database, with the most recent search conducted on September 13, 2023. The search query used was: `(basic reproduction number[TIAB] OR basic reproductive number[TIAB] OR basic reproduction ratio[TIAB] OR basic reproductive rate[TIAB] OR R0[TIAB]) NOT (R0 resection OR cancer)`, targeting papers with any synonyms of $R0$ in the title or abstract. This yielded 7,127 articles. We leveraged the ORKG-FLAN-T5 R0 LLM [39] to filter articles that did not report an $R0$ value as unanswerable; or otherwise provide structured JSON descriptions for articles with $R0$ estimates.

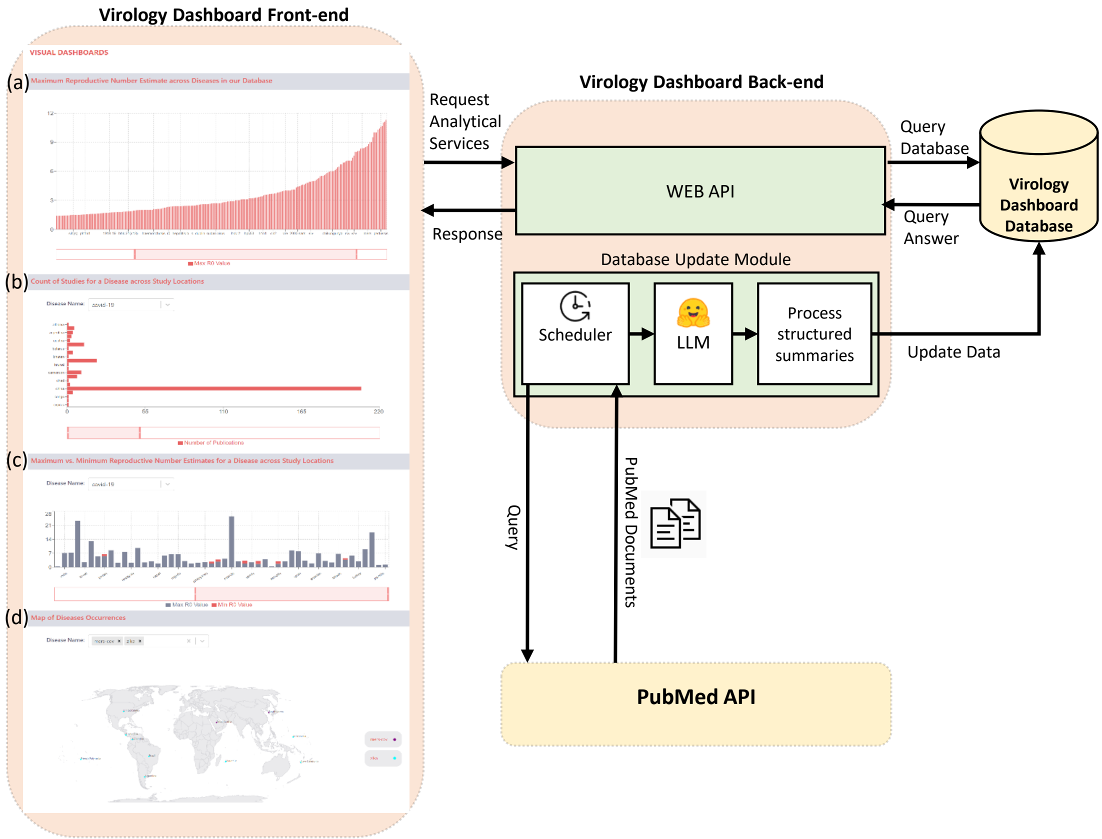

Figure 1: (Left image) A visual analytical dashboard in our next-generation information retrieval (IR) platform provides charts (a), (b), (c), (d) in a dashboard to help researchers make informed article filtering decisions. **(Right image)** The backend workflow, managed by a web API, handles database interactions for frontend rendering. It incorporates a schedule for database updates programmed to run monthly, with LLM queries supplying structured scientific knowledge before each update.

After filtering out unanswerables, 2,051 articles remained, yielding 2,736 structured summaries. The processed data was imported to a PostgreSQL 16 database, serving as the backend storage. The top 20 most represented infectious diseases in our initial database is shown in Table 1. Notably, the LLM's high precision confirmed that the top reported diseases are indeed ascertained infectious diseases. Our database covers studies from all seven continents.

2.3. The Information Retrieval (IR) Platform Workflow

The platform is accessible as a web application at the following URL: https://orkg.org/usecases/r0-estimates. The visualization dashboard widget and underlying workflow are displayed in Figure 1. In this workflow, the frontend communicates with the backend through a Web API for database queries and data retrieval. A Python script scheduler, programmed to run monthly, periodically updates the database with new articles querying PubMed and following the LLM processing cycle before updating the database with structured summaries. Our workflow maximizes the use of cutting-edge technology, including an optimized next-generation LLM.

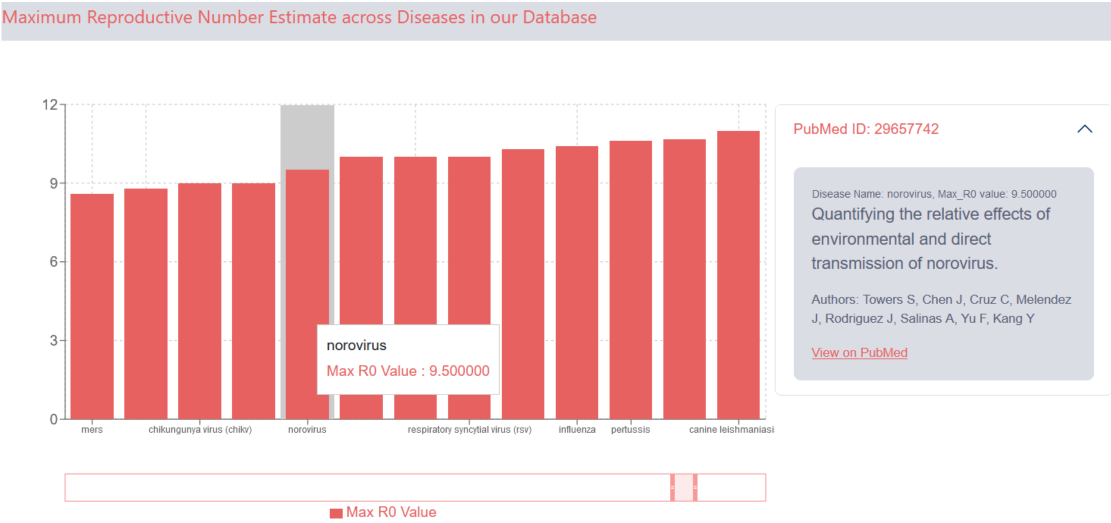

Figure 2: Chart (a) in Figure 1 displays maximum $R0$ values by disease to enhance scholarly publication filtering. The y-axis shows max $R0$ values, and the x-axis lists diseases. Users can filter by $R0$ range, and clicking a bar reveals underlying publication details with links to PubMed articles.

2.3.1. Charting the data: collating, summarizing, and reporting

Our IR platform includes three main components: 1) a statistics snapshot showing total papers, structured knowledge, infectious diseases, and locations, 2) a standard paper listing in a keyword-based table, filtered as needed, built with the ag-grid JavaScript library, and 3) a visual analytical dashboard with four charts addressing our research questions. This process involves *collating* relevant properties, selecting the best chart from the React chart library to *summarize* the response, and creating a query to *report* the visual summary. Each RQ is represented by a visual chart. E.g., **RQ1**, "What are the maximum $R0$ estimates reported for diseases in our database?" is illustrated with a bar chart that plots diseases on the x-axis against their maximum $R0$ values on the y-axis. Hovering over a bar displays the disease and its max $R0$. This interactive chart, which can be adjusted for specific $R0$ ranges, simplifies the comparison of $R0$ estimates across numerous studies. Clicking on a bar provides a direct link to the contributing article on PubMed, thereby enhancing scholarly information retrieval significantly beyond traditional methods.

3. Conclusion

In this poster paper, we present a POC for a new scholarly IR engine that enhances access and reduces the cognitive load of traditional, keyword-based searches. We address the inefficiencies of manual paper filtering in traditional IR systems, exacerbated by rapidly increasing publication volumes. Our approach models key research aspects for machine processing, paving the way for next-generation visual assistants that streamline scholarly research access.

References

[1] S. Fortunato, C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, et al., Science of science, Science 359 (2018) eaao0185.

[2] L. Bornmann, R. Haunschild, R. Mutz, Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases, Humanities and Social Sciences Communications 8 (2021) 1–15.

[3] L. Ermakova, E. SanJuan, S. Huet, H. Azarbonyad, G. M. Di Nunzio, F. Vezzani, J. D'Souza, S. Kabongo, H. B. Giglou, Y. Zhang, S. Auer, J. Kamps, Clef 2024 simpletext track: Improving access to scientific texts for everyone, Springer-Verlag, Berlin, Heidelberg, 2024, p. 28–35. URL: https://doi.org/10.1007/978-3-031-56072-9_4. doi:10.1007/978-3-031-56072-9_4.

[4] P. Fontelo, A. Gavino, R. F. Sarmiento, Comparing data accuracy between structured abstracts and full-text journal articles: implications in their use for informing clinical decisions, BMJ Evidence-Based Medicine 18 (2013) 207–211.

[5] W. Ammar, D. Groeneveld, C. Bhagavatula, I. Beltagy, M. Crawford, D. Downey, J. Dunkelberger, A. Elgohary, S. Feldman, V. Ha, et al., Construction of the literature graph in semantic scholar, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers), 2018, pp. 84–91.

[6] D. Pride, M. Cancellieri, P. Knoth, Core-gpt: Combining open access research and large language models for credible, trustworthy question answering, in: International Conference on Theory and Practice of Digital Libraries, Springer, 2023, pp. 146–159.

[7] S. Auer, A. Oelen, M. Haris, M. Stocker, J. D'Souza, K. E. Farfar, L. Vogt, M. Prinz, V. Wiens, M. Y. Jaradeh, Improving access to scientific literature with knowledge graphs, Bibliothek Forschung und Praxis 44 (2020) 516–529.

[8] A. Oelen, M. Stocker, S. Auer, Smartreviews: towards human-and machine-actionable reviews, in: Linking Theory and Practice of Digital Libraries: 25th International Conference on Theory and Practice of Digital Libraries, TPDL 2021, Virtual Event, September 13–17, 2021, Proceedings 25, Springer, 2021, pp. 181–186.

[9] A. Oelen, M. Y. Jaradeh, K. E. Farfar, M. Stocker, S. Auer, Comparing research contributions in a scholarly knowledge graph, in: CEUR workshop proceedings; 2526, volume 2526, Aachen: RWTH Aachen, 2019, pp. 21–26.

[10] H. Santos, V. Dantas, V. Furtado, P. Pinheiro, D. L. McGuinness, From data to city indicators: A knowledge graph for supporting automatic generation of dashboards, in: The Semantic Web: 14th International Conference, ESWC 2017, Portorož, Slovenia, May 28–June 1, 2017, Proceedings, Part II 14, Springer, 2017, pp. 94–108.

[11] T. Khodaveisi, H. Dehdarirad, H. Bouraghi, A. Mohammadpour, F. Sajadi, M. Hosseiniravandi, Characteristics and specifications of dashboards developed for the covid-19 pandemic: a scoping review, Journal of Public Health (2023) 1–22.

[12] O. Lezhnina, G. Kismihók, M. Prinz, M. Stocker, S. Auer, A scholarly knowledge graph-powered dashboard: Implementation and user evaluation, Frontiers in Research Metrics and Analytics 7 (2022) 934930.

[13] H. Santos, V. Dantas, V. Furtado, P. Pinheiro, D. L. McGuinness, From data to city indicators: A knowledge graph for supporting automatic generation of dashboards, in: The Semantic

Web: 14th International Conference, ESWC 2017, Portorož, Slovenia, May 28–June 1, 2017, Proceedings, Part II 14, Springer, 2017, pp. 94–108.

[14] M. Salehi, M. Arashi, A. Bekker, J. Ferreira, D.-G. Chen, F. Esmaeili, M. Frances, A synergetic r-shiny portal for modeling and tracking of covid-19 data, Frontiers in public health 8 (2021) 623624.

[15] D.-H. Yang, T.-W. Chien, Y.-T. Yeh, T.-Y. Yang, W. Chou, J.-K. Lin, Using the absolute advantage coefficient (aac) to measure the strength of damage hit by covid-19 in india on a growth-share matrix, European Journal of Medical Research 26 (2021) 1–11.

[16] Z. Zhu, K. Meng, J. Caraballo, I. Jaradat, X. Shi, Z. Zhang, F. Akrami, H. Liao, F. Arslan, D. Jimenez, et al., A dashboard for mitigating the covid-19 misinfodemic, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, 2021.

[17] D. Aristizábal-Torres, C. A. Peñuela-Meneses, A. M. Barrera-Rodríguez, An interactive web-based dashboard to track covid-19 in colombia. case study: five main cities, Revista de Salud Pública 22 (2023) 214–219.

[18] L. E. Hodgson, T. Leckie, A. Hunter, N. Prinsloo, R. Venn, L. Forni, Covid-19 recognition and digital risk stratification, Future Healthcare Journal 7 (2020) e47.

[19] R. Ravinder, S. Singh, S. Bishnoi, A. Jan, A. Sharma, H. Kodamana, N. A. Krishnan, An adaptive, interacting, cluster-based model for predicting the transmission dynamics of covid-19, Heliyon 6 (2020).

[20] J. P. Ulahannan, N. Narayanan, N. Thalhath, P. Prabhakaran, S. Chaliyeduth, S. P. Suresh, M. Mohammed, E. Rajeevan, S. Joseph, A. Balakrishnan, et al., A citizen science initiative for open data and visualization of covid-19 outbreak in kerala, india, Journal of the American Medical Informatics Association 27 (2020) 1913–1920.

[21] B. D. Wissel, P. Van Camp, M. Kouril, C. Weis, T. A. Glauser, P. S. White, I. S. Kohane, J. W. Dexheimer, An interactive online dashboard for tracking covid-19 in us counties, cities, and states in real time, Journal of the American Medical Informatics Association 27 (2020) 1121–1125.

[22] A. S. Peddireddy, D. Xie, P. Patil, M. L. Wilson, D. Machi, S. Venkatramanan, B. Klahn, P. Porebski, P. Bhattacharya, S. Dumbre, et al., From 5vs to 6cs: Operationalizing epidemic data management with covid-19 surveillance, in: 2020 IEEE International Conference on Big Data (Big Data), IEEE, 2020, pp. 1380–1387.

[23] Y. S. Bae, K. H. Kim, S. W. Choi, T. Ko, C. W. Jeong, B. Cho, M. S. Kim, E. Kang, Information technology–based management of clinically healthy covid-19 patients: lessons from a living and treatment support center operated by seoul national university hospital, Journal of medical Internet research 22 (2020) e19938.

[24] H. Florez, S. Singh, Online dashboard and data analysis approach for assessing covid-19 case and death data, F1000Research 9 (2020).

[25] I. Pathak, Y. Choi, D. Jiao, D. Yeung, L. Liu, Racial-ethnic disparities in case fatality ratio narrowed after age standardization: A call for race-ethnicity-specific age distributions in state covid-19 data, MedRxiv (2020).

[26] H. Ibrahim, S. Sorrell, S. C. Nair, A. Al Romaithi, S. Al Mazrouei, A. Kamour, Rapid development and utilization of a clinical intelligence dashboard for frontline clinicians to optimize critical resources during covid-19, Acta Informatica Medica 28 (2020) 209.

[27] A. Chande, S. Lee, M. Harris, Q. Nguyen, S. J. Beckett, T. Hilley, C. Andris, J. S. Weitz, Real-time, interactive website for us-county-level covid-19 event risk assessment, Nature human behaviour 4 (2020) 1313–1319.

[28] V. Marivate, H. M. Combrink, Use of available data to inform the covid-19 outbreak in south africa: a case study, arXiv preprint arXiv:2004.04813 (2020).

[29] A. Hohl, E. M. Delmelle, M. R. Desjardins, Y. Lan, Daily surveillance of covid-19 using the prospective space-time scan statistic in the united states, Spatial and spatio-temporal epidemiology 34 (2020) 100354.

[30] R. Carroll, C. R. Prentice, Using spatial and temporal modeling to visualize the effects of us state issued stay at home orders on covid-19, Scientific Reports 11 (2021) 13939.

[31] N. Marques da Costa, N. Mileu, A. Alves, Dashboard comprime_compri_mov: Multiscalar spatio-temporal monitoring of the covid-19 pandemic in portugal, Future Internet 13 (2021) 45.

[32] M. Hyman, C. Mark, A. Imteaj, H. Ghiaie, S. Rezapour, A. M. Sadri, M. H. Amini, Data analytics to evaluate the impact of infectious disease on economy: Case study of covid-19 pandemic, Patterns 2 (2021).

[33] F. Clement, A. Kaur, M. Sedghi, D. Krishnaswamy, K. Punithakumar, Interactive data driven visualization for covid-19 with trends, analytics and forecasting, in: 2020 24th International Conference Information Visualisation (IV), IEEE, 2020, pp. 593–598.

[34] B. E. Dixon, S. J. Grannis, C. McAndrews, A. A. Broyles, W. Mikels-Carrasco, A. Wiensch, J. L. Williams, U. Tachinardi, P. J. Embi, Leveraging data visualization and a statewide health information exchange to support covid-19 surveillance and response: application of public health informatics, Journal of the American Medical Informatics Association 28 (2021) 1363–1373.

[35] R. Arias-Carrasco, J. Giddaluru, L. E. Cardozo, F. Martins, V. Maracaja-Coutinho, H. I. Nakaya, Outbreak: a user-friendly georeferencing online tool for disease surveillance, Biological Research 54 (2021) 1–6.

[36] R. Chauhan, P. Goel, V. Kumar, N. Soni, et al., Understanding covid-19 using data visualization, in: 2021 international conference on advance computing and innovative technologies in engineering (ICACITE), IEEE, 2021, pp. 555–559.

[37] A. Oelen, J. D'Souza, M. Stocker, L. Vogt, K. E. Farfar, M. Haris, K. Fadel, M. Y. Jaradeh, V. Wiens, Covid-19 reproductive number estimates, 2020. URL: https://www.orkg.org/orkg/comparison/R44930. doi:10.48366/R44930.

[38] L. Gordis, Epidemiology e-book, Elsevier Health Sciences, 2013.

[39] M. Shamsabadi, J. D'Souza, S. Auer, Large Language Models for Scientific Information Extraction: An Empirical Study for Virology, in: Y. Graham, M. Purver (Eds.), Findings of the Association for Computational Linguistics: EACL 2024, Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 374–392. URL: https://aclanthology.org/2024.findings-eacl.26.

[40] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, The Journal of Machine Learning Research 21 (2020) 5485–5551.

[41] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le, Finetuned language models are zero-shot learners, arXiv preprint arXiv:2109.01652 (2021).

[42] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, et al., Scaling instruction-finetuned language models, arXiv preprint arXiv:2210.11416 (2022).
[43] S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, Y. Tay, D. Zhou, Q. V. Le, B. Zoph, J. Wei, et al., The flan collection: Designing data and methods for effective instruction tuning, arXiv preprint arXiv:2301.13688 (2023).