ECAI 2025
I. Lynce et al. (Eds.)

© 2025 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA251444

ADA-X: An Online System for Fully Automated, Explainable Review Aggregation

Lingjun Gao[a], Hafizh Muyassar[a], Yeva Hunanyan[a], Shane Pongpanich[a], Antonio Rago[a,b,*] and Francesca Toni[a]

[a]Department of Computing, Imperial College London, UK
[b]Department of Informatics, King's College London, UK

Abstract. Today's online platforms, e.g. in e-commerce, often offer users numerous competing options in single product categories, e.g. televisions or watches, making it difficult for the users to identify the best option to suit their preferences. To ease this process, many platforms provide users with simple scores resulting from the aggregation of other users' reviews. However, these scoring systems may oversimplify the underlying information and lack explanatory context. Our main contribution in this demonstration paper is a novel online system for aggregating customer reviews and explaining the aggregation to users.The system operates through a multi-stage pipeline: it first applies novel automatic ontology extraction methods using BERT or Large Language Models to identify key aspects from customer reviews, then constructs support and attack relations between these aspects using Argumentative Dialogical Agents (ADAs), an existing methodology for generating argumentative analyses of aspects. Finally, it generates ontology-driven, explainable aggregations of the reviews. We evaluate the performance of our system (which we call ADA-X) on the Amazon and Disneyland review datasets, focusing the ontology quality using the LLM-as-a-judge method and aggregation performance against the original Amazon and Disneyland ratings. The demonstration is available at https://ada-x.co.uk/.

1 Introduction

In today's era of material abundance, users are presented with an overwhelming number of products across various categories. Reviews made by existing buyers are often treated as a reliable source of information that can help users make their decisions. However, due to the sheer volume of reviews, it is unrealistic and not cognitively manageable for users to read them all. To simplify this process, platforms often aggregate reviews into a single rating, such as the Amazon rating score[1] in the e-commerce domain and the Tomatometer Score[2] (TS) in the movie domain. These scoring systems face two main challenges. First, they oversimplify the evaluation process. Products across different domains are rated highly or lowly by users based on their features, and a single scoring system cannot adequately represent the positive or negative contributions from each. The TS, for example, has been criticised for negatively impacting the movie industry.[3] Second, many of these scoring systems are based on machine learning algorithms that lack explainability. For example, the "Amazon's Choice" label, which is based on the Amazon review score, has been criticised by various sources for failing to provide insightful suggestions to consumers[4]. Some attempts have been made to use Large Language Models (LLMs) in an end-to-end manner, e.g. Amazon, but there is no way of guaranteeing that the aggregations or explanations from LLMs are faithful [21].

Argumentative Dialogical Agents (ADAs) [18] address this problem by utilising Natural Language Processing (NLP) and computational argumentation (see [2, 3] for overviews) techniques to provide explainable review aggregations. Indeed, argumentation is attracting attention in the area given its strengths in reasoning with conflicting knowledge [1, 8, 16, 23]. An ADA is based on argumentative interpretations of reviews and hand-crafted [10] or mined [17] ontologies, ensuring that the explanations are faithful to the review aggregations. A limitation of this approach when supported by hand-crafted ontologies is its reliance on the manual construction of features, which requires domain-specific knowledge [11, 20]. Meanwhile, the automated methods supported by mined ontologies originally utilised BERT [12], showing encouraging performance on Amazon product datasets [17]. Motivated by the strong performance of LLMs in logical reasoning tasks [6, 15, 24], another method was recently introduced for automatic ontology extraction via LLMs [22], though this has not been formally evaluated within the ADA context.

These recent developments inspired us to integrate both mined ontology approaches, i.e. the BERT- and LLM-based methods, to create a fully automated, explainable review aggregation system, which we call ADA-X. Our system, available as a website[5] or a codebase[6], includes a live demonstration that allows developers to explore the ontology constructed by both mining methods, as well as the information for supporting potential interactions between users and ADA-X. The general workflow of the proposed approach is illustrated in Figure 1. In addition, we evaluate the enhanced framework across multiple domains. The quality of the generated ontologies is assessed using the LLM-as-a-judge method [25], and the review aggregation results are compared against conventional methods for aggregating reviews, i.e. the Amazon customer ratings and the Disney rating. These experiments show that both ontology mining methods extracted appropriate terms to refer to the product and identify reasonable relations between aspects. Meanwhile, both methods achieve

* Corresponding Author. Email: antonio.rago@kcl.ac.uk.
[1] https://www.amazon.co.uk/
[2] https://www.rottentomatoes.com/
[3] https://www.theatlantic.com/amp/article/543090/
[4] https://www.wired.com/story/what-does-amazons-choice-mean/
[5] https://ada-x.co.uk/
[6] https://github.com/ImperialOntology/llm-review-aggregation

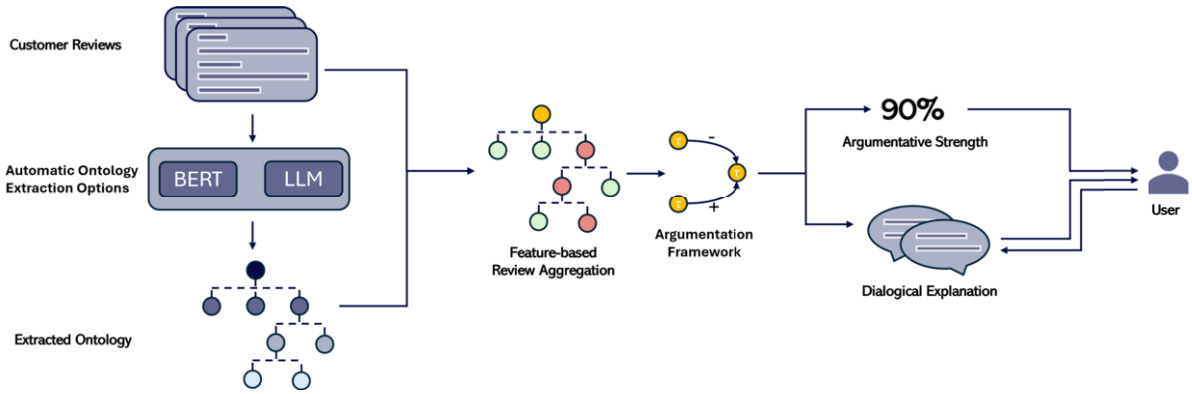


Figure 1. General workflow of the ADA-X system. We point readers to [17, 18, 22] for the technical details of each individual step.

similar aggregated scores compared with the original ratings from Amazon and Disneyland, indicating reliable performance of ADA-X. This work thus provides developers with an overview of how ADA-X can support fully automated, explainable review aggregation.

2 Functionalities

This section outlines the user interaction process with our online implementation of ADA-X, outlining its main functionalities.[7]

2.1 Dataset

We used two existing datasets: the Amazon product review dataset[8] and the Disneyland product reviews dataset[9]. The Amazon dataset was selected for its diversity of product categories, making it well-suited for evaluating the generalisability of the ontology extraction and review aggregation pipeline across a wide rage of products and user opinions. Due to the large number of Amazon customer reviews (over 500 million), five out of 33 domains were selected and 100,000 reviews were randomly sampled from each domain. The Disneyland reviews dataset was included to assess the ADA system's performance on a dataset where each product accumulates a significantly higher volume of reviews. Specifically, the dataset contains approximately 40,000 reviews distributed across only three Disneyland branches, in contrast to the Amazon Video Games category, where 100,000 reviews span roughly 17,000 distinct products.

The Analysis page, see Figure 2, allows users to configure their analysis parameters based on the selected data source. It begins by fetching and displaying the available product categories associated with the chosen data source, allowing the user to select a specific category. Once a category is selected, the app dynamically retrieves the two automatic ontology extraction methods, i.e. the BERT-based [17] and LLM-based [22] methods.

2.2 Automatic Ontology Extraction

There is a significant distinction between the BERT- and LLM-based approaches. The BERT-based method relies on a manually

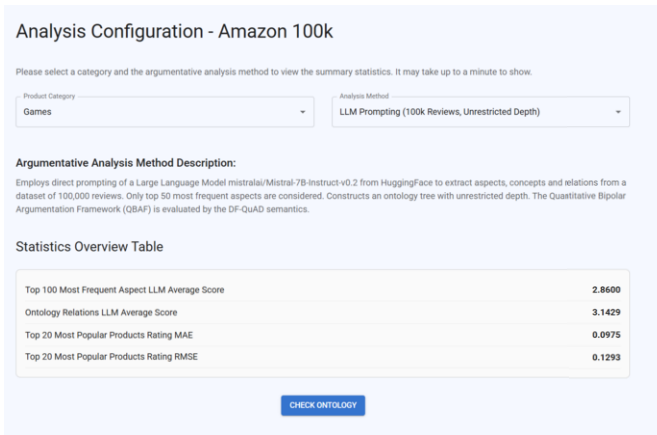


Figure 2. Analysis page of the online ADA-X system, allowing users select the product category and analysis method.

constructed dataset to fine-tune the model for ontology extraction, while the LLM-based method prompts the Mistral-7B-Instruct-v0.2 model [14] directly without additional training.

Upon selecting both a category and an ontology extraction method, a brief description of the selected method is shown, alongside an overview table displaying four performance metrics: the top 100 most frequent aspects LLM-as-a-judge score, the ontology relations LLM-as-a-judge score, the mean absolute error (MAE), and the root mean square error (RMSE) of product argument strength relative to average user rating. After configuring these options, users can proceed to the Review Aggregation page by clicking the "Check Ontology" button.

2.3 Argumentative Dialogical Agent

The Review Aggregation page serves as the core visualisation and exploration interface of the web application. Initially, it fetches the top 10 products from the selected category and allows the user to choose one for ontology inspection. Based on the chosen product and ontology extraction method, the app presents the product level statistics as well as the ontology tree of the corresponding product.

The product-level statistics provide a comparison between the conventional methods for producing the aggregated scores, i.e. Amazon ratings and Disneyland ratings in this paper,and the ontology product score based on the ADA framework. A closer value between these two figures demonstrates the effectiveness of the ontology extraction

[7] Due to the computational cost of running real-time analysis, the ADA system on our website uses pre-generated outputs. These results were fully automatically generated without any manual modifications. This approach ensures that users can experience the system's capabilities as they would appear in real-world deployment while maintaining reproducibility and explainability of its outputs.

[8] https://huggingface.co/datasets/McAuley-Lab/Amazon-Reviews-2023

[9] https://www.kaggle.com/datasets/arushchillar/disneyland-reviews

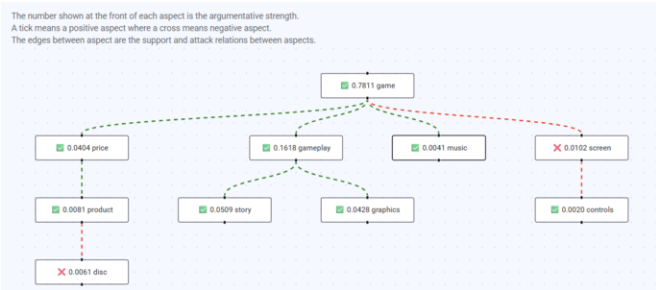


Figure 3. Review Aggregation page showing the argumentation framework constructed for a product in the game category. The colours of the edges between aspects represent the positive support (green) and negative attack (red) relations between aspects, while the ticks and crosses represent the aspects' positive or negative, respectively, overall contribution to the product.

and review aggregation process. Finally, shown beneath the statistics is the argumentation framework representing the review aggregation, as shown in Figure 3. Each aspect is assigned an argumentative *strength*, which is calculated by an argumentative semantics (see [18]). The argumentative strength of the product then corresponds to its score in the review aggregation. A prune tree option has been provided on the top of review aggregation page, which allows the user to filter out the aspects with zero strength shown in the argumentation framework (by most argumentative semantics, zero strength arguments are irrelevant [4]).

Beneath the ontology tree, an interactive dialogue-agent was implemented. This allows users to engage with predefined queries, such as "Why was the aspect good?", facilitating detailed exploration of explanations and user sentiments related to specific aspects. These explanations are dynamically generated by analysing aspect polarity, supporting and attacking features, argumentative strengths and related phrases within the ontology.

3 Empirical Evaluation

In this section, we detail our evaluation of the performance of two major components of ADA-X: the ontology extraction and the review aggregation. For the former, we evaluate the generated ontology quality by using the LLM-as-a-judge method [25], while the latter has been evaluated using the comparison between the argumentative strength and the original score by conventional review aggregation methods. Table 1 summarises the results.

3.1 Evaluation of the Ontology Extraction

To assess the quality of the generated ontology, specifically the extracted aspects and relationships, we employed the LLM-as-a-judge approach [25] using Gemini 1.5 Flash [19]. Scores range from 1 to 5, where a higher score indicates better performance. We evaluated two components of the ontology extraction: its aspects and its relations. For aspect evaluation, the score is calculated as the average rating assigned by the LLM to the top 100 most frequently occurring aspects in the review corpus. The evaluation criteria include relevance, specificity, clarity, and product fit. For relation evaluation, the score is computed as the average LLM-as-a-judge score across all relations in the constructed ontology tree, with the judge assessing each based on logical hierarchy, contextual fit, and clarity (technical details of this evaluation method can be found in [5]).

Based on the results in Table 1, the BERT-based method generally performs better in aspect extraction than the LLM-based method,

Category	Ontology Extraction				Review Aggregation			
	Aspects ↑		Relations ↑		MAE ↓		RMSE ↓	
	LLM	BERT	LLM	BERT	LLM	BERT	LLM	BERT
Mixers	2.74	**3.30**	**2.83**	2.56	**0.071**	0.073	**0.084**	0.096
Necklaces	2.57	**3.23**	1.40	**1.81**	0.104	**0.097**	0.114	**0.109**
Games	**2.86**	2.75	**3.14**	3.11	**0.098**	0.127	**0.129**	0.159
Television	3.22	**3.44**	3.22	**3.35**	**0.140**	0.165	**0.145**	0.170
Watches	3.16	**3.57**	**2.93**	2.38	0.087	**0.086**	0.110	**0.098**
Disney	**3.34**	3.25	**3.00**	2.50	0.221	**0.205**	0.222	**0.210**
Average	2.98	**3.26**	**2.75**	2.62	**0.120**	0.126	**0.134**	0.140

Table 1. LLM-as-a-judge scores in the evaluation of ontology extraction methods; MAE and RMSE in the evaluation of review aggregation methods.

with an average rating of 3.26. However, this trend is reversed in relation construction, where the LLM-based model shows superior performance and an average rating of 2.75. We thus posit that the specific contextual knowledge in pre-trained BERT models may be advantageous in entity recognition tasks [9], while LLMs may have an advantage in detecting relations between entities, as has been found in relation-based argument mining [7, 13]. We leave deeper validation of these hypotheses to future work. Overall, these scores provide a quantitative assessment of the model outputs, highlighting the strengths and weaknesses of each method in extracting meaningful semantic information from customer reviews.

3.2 Evaluation of the Review Aggregation

To evaluate the performance of the review aggregation, the MAE and RMSE were calculated between the product average rating and the ontology product score in both the Amazon and Disney datasets.

Given that the scale of both MAE and RMSE ranges was from 0 to 1, both methods gave promising averages of less than 0.140 as shown in Table 1. This indicates a reasonable alignment between the aggregated argument strengths and the overall user sentiment. For the Amazon product categories, both the LLM-based and BERT-based methods demonstrate comparable performance. The Disney dataset, despite its relatively high number of reviews per product, shows the highest MAE and RMSE values for both methods, which may be attributed to the unique characteristics of the reviews or the sentiment expressions within this dataset.

4 Conclusions and Future Work

This paper successfully implements the ADA-X system with the integration of two distinctive automatic ontology extraction method with the ADA framework. Compared with traditional review aggregation approaches, ADA-X offers more explainable insights for customers, potentially helping them better understand product attributes and performance. We demonstrated the system's effectiveness using two datasets: Amazon product reviews and Disneyland reviews. ADA-X achieved promising scores when compared with the aggregated ratings provided by the respective platforms.

This work opens up a number of avenues for future work. For example, the modular design of the codebase enables future developers to incorporate new features and experiment with alternative ontology extraction methods for further research and improvement. Moreover, it would be interesting to undertake further experimentation in datasets across different product types, e.g. in streamed products such as movies or music, as well as user studies in real world settings, which will be crucial for evaluating ADA-X's effectiveness along human-centric metrics such as comprehensibility or trust.

References

[1] X. F. Aran and E. Plaza. Concept discovery and argument bundles in the experience web. In *ICCBR*, pages 108–123, 2016. doi: 10.1007/978-3-319-47096-2_8. URL https://doi.org/10.1007/978-3-319-47096-2_8.

[2] K. Atkinson, P. Baroni, M. Giacomin, A. Hunter, H. Prakken, C. Reed, G. R. Simari, M. Thimm, and S. Villata. Towards artificial argumentation. *AI Magazine*, 38(3):25–36, 2017.

[3] P. Baroni, D. Gabbay, M. Giacomin, and L. van der Torre, editors. *Handbook of Formal Argumentation*. College Publications, 2018.

[4] P. Baroni, A. Rago, and F. Toni. From fine-grained properties to broad principles for gradual argumentation: A principled spectrum. *Int. J. Approx. Reason.*, 105:252–286, 2019. doi: 10.1016/J.IJAR.2018.11.019. URL https://doi.org/10.1016/j.ijar.2018.11.019.

[5] C. Brewster. Ontology learning from text: Methods, evaluation and applications. *Comput. Linguistics*, 32(4):569–572, 2006. doi: 10.1162/COLI.2006.32.4.569. URL https://doi.org/10.1162/coli.2006.32.4.569.

[6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In *NeurIPS*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[7] J. Cabessa, H. Hernault, and U. Mushtaq. Argument mining with fine-tuned large language models. In *COLING*, pages 6624–6635, 2025. URL https://aclanthology.org/2025.coling-main.442/.

[8] D. Ceolin, G. Primiero, M. Soprano, and J. Wielemaker. Transparent assessment of information quality of online reviews using formal argumentation theory. *Inf. Syst.*, 110:102107, 2022. doi: 10.1016/j.is.2022.102107. URL https://doi.org/10.1016/j.is.2022.102107.

[9] L. Chen and G. Varoquaux. What is the role of small models in the LLM era: A survey. *CoRR*, abs/2409.06857, 2024. doi: 10.48550/ARXIV.2409.06857. URL https://doi.org/10.48550/arXiv.2409.06857.

[10] O. Cocarascu, A. Rago, and F. Toni. Extracting dialogical explanations for review aggregations with argumentative dialogical agents. In *AAMAS*, pages 1261–1269, 2019. URL http://dl.acm.org/citation.cfm?id=3331830.

[11] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. P. Kuksa. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537, 2011. doi: 10.5555/1953048.2078186. URL https://dl.acm.org/doi/10.5555/1953048.2078186.

[12] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019. doi: 10.18653/V1/N19-1423. URL https://doi.org/10.18653/v1/n19-1423.

[13] D. Gorur, A. Rago, and F. Toni. Can large language models perform relation-based argument mining? In *COLING*, pages 8518–8534, 2025. URL https://aclanthology.org/2025.coling-main.569/.

[14] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

[15] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. In *NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html.

[16] J. Mumford, S. Sarkadi, K. Atkinson, and T. J. M. Bench-Capon. Applying argument schemes for simulating online review platforms. In *COMMA*, pages 169–180, 2024. doi: 10.3233/FAIA240319. URL https://doi.org/10.3233/FAIA240319.

[17] J. Oksanen, O. Cocarascu, and F. Toni. Automatic product ontology extraction from textual reviews. *CoRR*, abs/2105.10966, 2021. URL https://arxiv.org/abs/2105.10966.

[18] A. Rago, O. Cocarascu, J. Oksanen, and F. Toni. Argumentative review aggregation and dialogical explanations. *Artif. Intell.*, 340:104291, 2025. doi: 10.1016/J.ARTINT.2025.104291. URL https://doi.org/10.1016/j.artint.2025.104291.

[19] M. Reid, N. Savinov, D. Teplyashin, D. Lepikhin, T. P. Lillicrap, J.-B. Alayrac, R. Soricut, A. Lazaridou, O. Firat, J. Schrittwieser, I. Antonoglou, R. Anil, S. Borgeaud, A. M. Dai, K. Millican, E. Dyer, M. Glaese, T. Sottiaux, B. Lee, F. Viola, M. Reynolds, Y. Xu, J. Molloy, J. Chen, M. Isard, P. Barham, T. Hennigan, R. McIlroy, M. Johnson, J. Schalkwyk, E. Collins, E. Rutherford, E. Moreira, K. Ayoub, M. Goel, C. Meyer, G. Thornton, Z. Yang, H. Michalewski, Z. Abbas, N. Schucher, A. Anand, R. Ives, J. Keeling, K. Lenc, S. Haykal, S. Shakeri, P. Shyam, A. Chowdhery, R. Ring, S. Spencer, E. Sezener, and et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530, 2024. URL https://doi.org/10.48550/arXiv.2403.05530.

[20] J. P. Turian, L. Ratinov, and Y. Bengio. Word representations: A simple and general method for semi-supervised learning. In *ACL*, pages 384–394, 2010. URL https://aclanthology.org/P10-1040/.

[21] M. Turpin, J. Michael, E. Perez, and S. R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *NeurIPS*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html.

[22] E. Wong. Automatic extraction of ontologies via large language models. Master's thesis, Impeiral College, 2024.

[23] A. Z. Wyner, J. Schneider, K. Atkinson, and T. J. M. Bench-Capon. Semi-automated argumentative analysis of online product reviews. In *Computational Models of Argument - Proceedings of COMMA 2012, Vienna, Austria, September 10-12, 2012*, volume 245, pages 43–50. IOS Press, 2012. doi: 10.3233/978-1-61499-111-3-43. URL https://doi.org/10.3233/978-1-61499-111-3-43.

[24] C. Xie, Y. Huang, C. Zhang, D. Yu, X. Chen, B. Y. Lin, B. Li, B. Ghazi, and R. Kumar. On memorization of large language models in logical reasoning. *CoRR*, abs/2410.23123, 2024. doi: 10.48550/ARXIV.2410.23123. URL https://doi.org/10.48550/arXiv.2410.23123.

[25] L. Zheng, W. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *NeurIPS*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.