

The Jaseci Programming Paradigm and Runtime Stack: Building Scale-out Production Applications Easy and Fast

Jason Mars, *Member, IEEE*, Yiping Kang, Roland Daynauth, Baichuan Li, Ashish Mahendra, Krisztian Flautner, Lingjia Tang, *Member, IEEE*

Abstract—Today’s production scale-out applications include many sub-application components, such as storage backends, logging infrastructure and AI models. These components have drastically different characteristics, are required to work in collaboration, and interface with each other as microservices. This leads to increasingly high complexity in developing, optimizing, configuring, and deploying scale-out applications, raising the barrier to entry for most individuals and small teams. We developed a novel co-designed runtime system, **Jaseci**, and programming language, **Jac**, which aims to reduce this complexity. The key design principle throughout Jaseci’s design is to raise the level of abstraction by moving as much of the scale-out data management, microservice componentization, and live update complexity into the runtime stack to be automated and optimized automatically. We use real-world AI applications to demonstrate Jaseci’s benefit for application performance and developer productivity.

Index Terms—Serverless Computing, Artificial intelligence, Warehouse-Scale Computing, Runtimes, Programming Languages.

1 INTRODUCTION

HERE has been a fundamental shift in how we build software over the last two decades. Complexity continues to rise and productivity declines for developers as it becomes increasingly difficult to build and maintain production-grade applications. This increased complexity is rooted in three challenge areas:

Challenge 1 - Application data has increased in scale, source, and diversity, making data management difficult for scale-out applications, especially where performance is concerned. Specifically, decisions on how to structure schemas, what data to be persistent, what data to be stored in distributed memcache, and how the various API’s and protocols to be designed bring significant complexity and poor maintainability as requirements change.

Challenge 2 - Significant complexity emerges in deciding which parts of the application should be separated out as standalone microservices vs be linked within the same address space. Today, these decisions are static and cannot be changed based on the runtime conditions of the application and the underlying cluster environment. Indeed, realizing optimal or near-optimal performance is not commonly achieved [1], and the relative performance of different design points changes over time.

Challenge 3 - Introducing (or removing) functionality often requires pushing a new version of an entire application or container to replace the old. The idea of replacing a single method or a single class of a running live service/application would be unthinkable for current scale-out applications. The complexity of updating scale-out software impedes the rate at which developers innovate.

Because of these complexities, the barrier to entry has grown for individuals and small teams to build scale-out applications as sophisticated skill sets with inter-role dependencies are required (Figure 1A).

In this work, we propose a runtime stack (**Jaseci**) and a programming model (with **Jac**) that significantly reduce the complexity of building and deploying applications. Jaseci and Jac introduce three abstractions at the programming language level that hides and automates the optimization of complex design decisions across our challenge areas.

1) A **graph-based data representation semantic** is realized in the Jac language’s core design that allows the developer

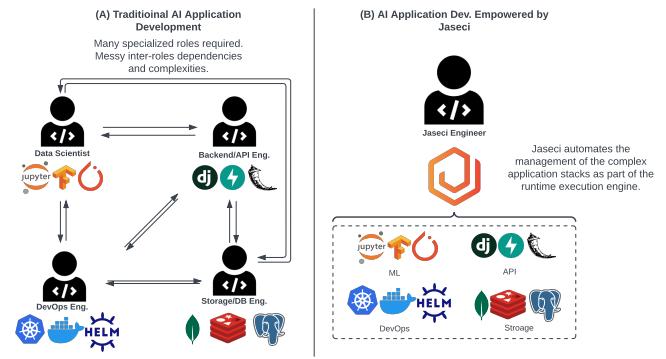


Fig. 1: Typical development team for a production-grade application without Jaseci (left) and with Jaseci (right).

to articulate compute and data coupling with nodes, edges, and graphs. We call this a **data-spatial programming model**. Jaseci then subsumes the responsibility of mapping data to the underlying distributed data memory and storage hierarchy. Using the data locality information inherently embedded in the structure of the graph and how scopes of code are accessed throughout execution, Jaseci automatically optimizes the distribution of data across the memory and data hierarchies.

2) A **control-flow abstraction, actions**, simplifies the transfer of control between components and hides whether the component code lives in the virtual address space of the running application or outside within a microservice (or FaaS). Developers declare and make action calls in their program to interact with functionality without having to explicitly decide or even know how they are *bound*. These bindings are controlled by Jaseci and can also change dynamically based on runtime monitoring.

3) The **compute/data encapsulation abstraction, walkers**, are introduced to facilitate and reduce the complexity of introducing new functionality to an existing application. Walkers enable dynamic injection and removal of functionality at fine granularity during application runtime as they silo functionality and data. These walkers execute independently on the graph through *walks* and communicate only by leaving data in the graph for other walkers to access. This approach further enables the data-spatial programming style of fusing compute with data and allows features and functionality to be independently and dynamically added (and removed) to a running scale-out application.

• The authors are with the University of Michigan. Email: {profmars, ypkang, daynauth, patrli, ashmahan, manowar, lingjia}@umich.edu.

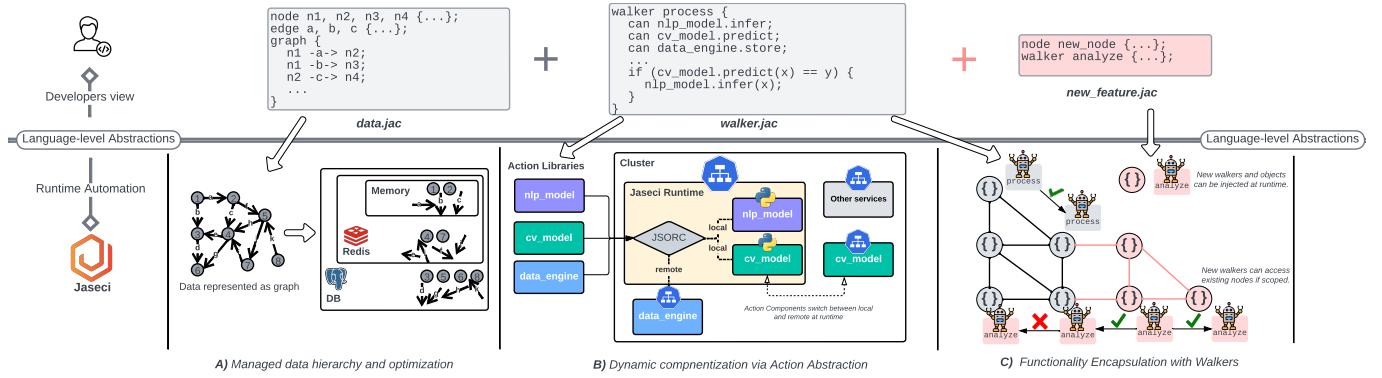


Fig. 2: Architecture of three key abstractions of Jaseci and Jac.

With these three key abstractions, Jaseci and Jac aim to 1) reduce the domain-specific special skill sets required and enable a small team or a single developer to build high-performance production-grade applications (Figure 1B), and 2) increase the rate at which developers can experiment and innovate new ideas at scale.

Jaseci and Jac are open-source [2], [3], [4], [5], fully functional, and have already been used by dozens of programmers for the creation of production software. Jaseci deployment already supports tens of thousands of production queries per day across six commercial products [6], [7], [8], [9], [10], [11]. In this work, we describe the key insights of these novel abstractions and show qualitatively and quantitatively how they help developers.

2 JASECI DESIGN AND IMPLEMENTATION

In this section, we describe the design of Jaseci (shown in Figure 2) with respect to the three challenges presented in Sec. 1.

2.1 Language-level Graph-based Data Representation

Jaseci uses a rich graph-based semantic as the key primitive at the language level for which both data and compute is articulated, as shown in Figure 2A. In the Jac language, graphs, nodes, and edges are first-class citizens. Developers interact with nodes and edges in the language (*data.jac* in Fig. 2), and they do not have to make decisions on data schema design, what should be persistent, and what should be cached. Jaseci runtime automates these decisions based on the graph semantics and the application behavior. (For brevity, more can be found about these abstractions in Jac's public documentation [12].) Under the language-level abstraction, Jaseci runtime manages nodes and edges across a three-level data hierarchy, including memory, a distributed memory caching layer, and a database for persistent storage. We are able to leverage the data locality information inherently represented in the graph structure and the emergent patterns of how walkers access node and edge objects. As a baseline, the LRU (Least Recently Used) memory caching policy for the Redis layer proves is used in production today to leverage this information, however this graph path information enables a new landscape of prefetching and is future work.

A technique we have designed and implemented in production is a smart graph object data packing approach we call *Fast Edges*. This approach is an adaptive node/edge fusion technique, further improves performance across the six production application workloads. We present initial performance measurements of Fast Edges in Section 4.1.

2.2 Dynamic Componentization via Actions

Figure 2B illustrates the *actions* abstraction and how it simplifies the process of application componentization for developers. When programming in Jac, developers declare action

calls with the *can* keyword in their code. How this action is bound to the application and where the action execution is going to occur is handled by Jaseci. As shown in Figure 2B, this example Jac program declares three action calls in *walker.jac*, *nlp_model*, *cv_model* and *data_engine*. Then at runtime, Jaseci configures *nlp_model* and *cv_model* as local libraries and *data_engine* as a remote microservice in the cluster, which are then bound to the corresponding action calls. The decision of whether an action should be bound as a locally linked library in the application's address space or bound as an RPC call to a remote microservice or FaaS service is abstracted away from the developer. Actions can also be used to interface existing software built in other languages and frameworks.

The key innovation of the *actions* abstraction is that it enables *late-binding* such that an automated orchestrator (JSORC in Fig. 2B) can serve as a subsystem to steer the decisioning of how an application should be componentized after the application is launched and potentially change those decisions dynamically. For example, in Fig. 2B, the action *cv_model* is switched from local library to remote microservice after the initial componentization decision. JSORC regularly analyzes the current runtime system state and the underlying hardware and devises and applies the best componentization decision. We describe in detail the design and algorithm of JSORC in Section 3.

2.3 Functionality Encapsulation with Walkers

In the current application development practice, introducing new features to a mature scale-out application requires the developers to apply the code changes to the code base, and push a new version of the entire application or container to replace the old version. Walkers provide the ability to encapsulate, inject and remove functionality at a fine granularity analogous to live adding a single method or function in a running application. However, a walker's scope is highly constrained to only accessing the state within itself and on the node/edge that it sits on during a *walk*. This approach provides a siloed guarantee that walkers can only communicate by writing and reading state spatially to the graph. We call this style of programming **data-spatial programming**, and introduces a new intuitive way of thinking about solving problems. For example, in Figure 2C, a new walker *analyze* is introduced to a running application, along with a new node type.

Additionally, this data-spatial approach makes available constraints and guarantees that reduce the risk of bugs and regression when introducing new functionality to software. For example, in the context of building RESTful API endpoints, a statically discernible scope of data that is accessible (visible) can be guaranteed throughout program execution. Developers can limit nodes/edges access to only certain walkers directly and intuitively in the language. In the example in Figure 2C, the *analyze* walker is not allowed to traverse

to certain existing nodes, whose scope limits access to only certain walkers (e.g. the existing *process* walker). We present a concrete example using walkers in a real-world production case study in Section 4.4.

3 DESIGN AND IMPLEMENTATION OF THE ACTION COMPONENTIZATION SUBSYSTEM

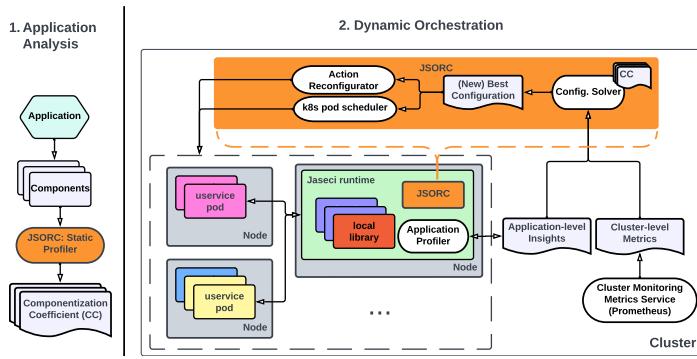


Fig. 3: Design of the Componentization Subsystem, JSORC

In this section, we describe the design of the Jaseci Orchestrator (JSORC), the subsystem that decides the action componentization at runtime (Fig. 2B). The key intuition of JSORC is that *the optimal componentization decision can be inferred from application-level runtime insights and cluster-level resources monitoring metrics*. JSORC retrieves these insights and metrics from other Jaseci subsystems and devises and applies the best componentization decision in real time. Figure 3 shows the overview of JSORC, which consists of two phases: application analysis (left) and dynamic orchestration (right).

Application Analysis - In order to accurately infer the ideal componentization configuration for the best QoS, it is imperative to first understand the components at play and their characteristics. Before deploying an application, JSORC profiles the actions required by the application. Specifically, we measure the latency of the action calls when configured as a remote microservice and a locally bound library, respectively, and define the ratio of remote vs. local latency as the Componentization Coefficient (CC). In addition, we profile the memory requirement of each action library.

Dynamic Orchestration - Many dynamic factors can influence the optimal componentization configuration, including changes in application request patterns, available hardware resources, and cost requirements. We design JSORC to dynamically adapt to these dynamic factors and periodically adjust its decision during runtime. JSORC harnesses the continuous profiling information provided by the Jaseci runtime application profiler and cluster-wide metrics services (Prometheus in Jaseci’s case). The application profiler provides application-level insights including action utilizations. Prometheus monitors the cluster and provides performance and utilization metrics of the deployed pods and nodes in the cluster. The Config. Solver uses these insights, metrics, and action profiles (e.g. CC) to infer a, potentially new, best componentization configuration. For new configurations, the Pod Scheduler creates new Kubernetes pods and/or removes existing ones, and the Reconfigurator updates the action call bindings.

4 EXPERIMENTS AND ANALYSIS

We perform a number of experiments and analyses with Jaseci and present our findings in this section.

4.1 Optimizing Data Performance with *Fast Edge*

With the introduction of *walkers*, *edges*, and *nodes* as first-order primitives in Jac, the way these data elements are stored introduces a new landscape of optimizations. We introduce one such optimization, *Fast Edges*, to optimize the read/write

performance of nodes and edges, especially for graphs with a large number of nodes. The key idea is that when a node is accessed, its connecting edges are often accessed next or soon, and as such, fusing edges together with nodes can reduce database access and improve data performance. Jaseci automatically marks an edge as a Fast Edge if its context data is smaller than a parametrized threshold. Normal edges are stored as separate objects in the database, while fast edges are fused and saved with their source and destination nodes. A Fast Edge is re-created in memory when its associated node is loaded.

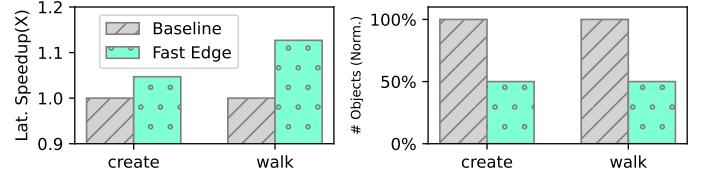


Fig. 4: Fast Edge

We use Myca [6], an in-production AI-powered personal productivity platform with 100s of users and 1000s of queries/day, to evaluate the efficacy of Fast Edge. Figure 4 shows, for two request types (create and walk), the latency speedup and reduction in Redis access achieved by Fast Edge. Fast Edge improves request latency by up to 13% and reduces # of objects accessed by up to 50.1%. Motivated by these promising results, we plan to design and implement the next evolution of Fast Edge, where the runtime selectively fuses nodes with adjacent nodes in addition to edges and combines access for subgraphs.

4.2 Characterizing the Componentization Problem

The decision of componentization has significant implications on application performance. We first characterize the difference in latency of the same action when configured as a remote vs. local component. Figure 5a compares the remote action latency of five state-of-the-art transformer-based AI models, including BERT [13], [14], [15] and USE [16]. We observe that **1**) remote-linked microservice actions incur longer execution latency than their local counterparts due to the additional network communication overhead and **2**) this latency degradation varies greatly depending on the component, due to difference in data to compute ratios.

We then characterize the distribution of performance profiles of the end-to-end application across the spectrum of different configurations of which actions are locally binded and which are remote microservices. We use an application with five AI model actions, which leads to 32 different component configurations. Figures 5b-d show the distribution of QPS, average latency, and 99th percentile latency as box plots grouped by the number of local components, where 0 *locals* means all five components are remote and 1 *locals* groups the performance of configs with one local component and four remote, etc. The results show a high impact of component configuration decisions on the overall performance of the application. We observe a large performance variation depending on which particular component is configured as local vs remote, as demonstrated by the large gap between the box boundaries/whiskers.

4.3 JSORC Dynamic Orchestration

There exists a large design space for beneficial heuristics for JSORC’s dynamic policy. In this experiment, we focus on a generalized policy that centered around optimizing for the best common case performance. In this policy, at runtime, JSORC regularly enters an evaluation phase where it tries out each possible configuration for a short time period and selects the one with the best performance as the applied configuration for the application until the next evaluation

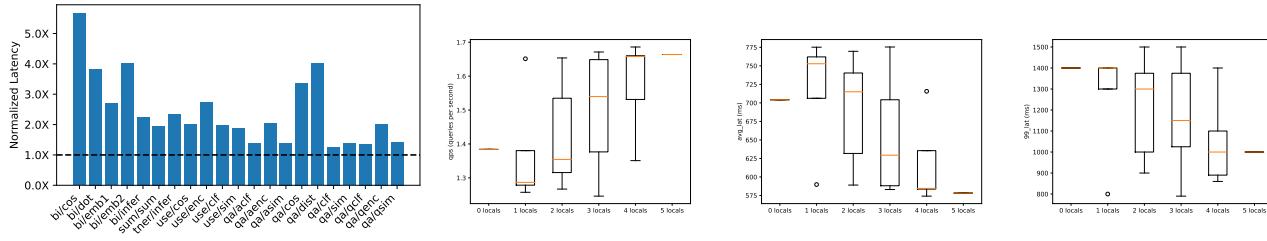


Fig. 5: From left to right: (a): Average request latency of each model action when configured as remote microservice, normalized to that of locally bound library. (b,c,d): Performance of microservice configurations with a varying ratio of local vs. remote modules.

phase is scheduled. The benefit of this evaluation-based policy is that it is agnostic to the heterogeneity of the cloud hardware resources and application characteristics.

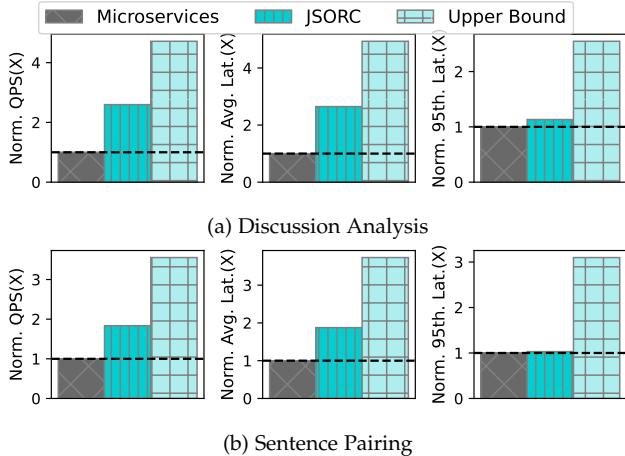


Fig. 6: Perf. with JSORC and remote microservice.

We conduct real-system experiments using two text analytic AI applications, *Discussion Analysis* (DA) analyzes open-ended group discussion transcript on social topics and *Sentence Pairing* (SP) pairs up related text from two disjoint lists. They each use two actions for two different AI models. Figure 6 shows the throughput, average, and tail latency of the application under JSORC's dynamic policy and all remote microservices. Results are normalized to all microservices configurations.

The current version of JSORC achieves up to $2.64 \times$ latency and $2.59 \times$ throughput improvement compared to the static config of all remote microservices. We show the Upper Bound performance, where all actions are configured as locally bound libraries in the Jaseci runtime process, and no communication over the network is required. This configuration leads to a single monolithic process with high memory consumption, requiring more costly hardware. Therefore, it is not always desired or feasible, especially at scale and high request load, and an adaptive and dynamic technique is needed. In addition, we observe a more significant speedup in throughput and average latency than tail latency. This is because of the additional time that is spent when JSORC is switching components between local and remote. We plan to introduce adaptive orchestration policies to reduce the frequency of component switching required to rein in the tail. Other promising heuristics we plan to explore include using application-level behavior tracking and cluster-level performance counters to predict the best componentization configuration without having to evaluate.

4.4 Case Study: Daily Summary Feature in Myca

We use a qualitative approach with an in-production application as a case study to demonstrate how a developer uses Walkers to nimbly introduce new functionality into running

applications. Myca [6], a personal productivity platform with hundreds of users, uses NLP AI models to provide users with insights into how they spend their time. A Myca developer introduced a new feature to summarize users' daily activity over a certain time period. The developer wrote a single walker *daily_summary* that traverses the user's graph from one *day* node to the next, collecting data from that day (e.g., completed tasks) and making action calls to a summarization component using the T5 transformer model [17]. The entire implementation of this feature requires ~ 25 lines of Jac code and is encapsulated in a single Walker, while typically such a feature would require hundreds of lines of code in a traditional programming model. After implementing the first prototype of the feature, the developer deployed it to the running production Jaseci instance by dynamically injecting the *daily_summary* walker, without touching the existing codebase or features. After beta testing with real users for two weeks, the developer decided to incorporate users' feedback into a v2 and removed the feature from production by just removing the *daily_summary* walker.

5 CONCLUSION

There is increasing complexity in developing and optimizing today's production scale-out applications. We introduce a novel co-designed runtime system *Jaseci* and programming language *Jac* to reduce this complexity and improve developer productivity. In this work, we focus on three key language-level abstractions that enable runtime automation for data management, microservice optimization, and functionality encapsulation. *Jaseci* and *Jac* are open-source and have been used to create six production software.

REFERENCES

- [1] S. Newman, *Monolith to microservices: evolutionary patterns to transform your monolith*. O'Reilly Media, 2019.
- [2] JaseciLabs, "Jaseci home," <https://jaseci.org/>, 2022, [Online; accessed 10-May-2022].
- [3] —, "The official jaseci code repository," <https://github.com/Jaseci-Labs/jaseci>, 2022, [Online; accessed 10-May-2022].
- [4] —, "Profile of jaseclabs pypi," <https://pypi.org/user/jaseclabs/>, 2022, [Online; accessed 10-May-2022].
- [5] —, "Jaseci's docker hub profile," <https://hub.docker.com/u/jaseci>, 2021, [Online; accessed 27-Jan-2023].
- [6] myca, "Myca.ai, growth via reflection," <https://myca.ai/>, 2022, [Online; accessed 10-May-2022].
- [7] HomeLendingPal, "Intelligent mortgage advisor - home lending pal," <https://www.homelendingpal.com/>, 2022, [Online; accessed 10-May-2022].
- [8] ZeroShotBot, "Next-gen ai, zeroshotbot," <https://zeroshotbot.com/>, 2022, [Online; accessed 10-May-2022].
- [9] TrueSelph, "Unleash your true selph, trueselph," <https://trueselph.com/>, 2022, [Online; accessed 10-May-2022].
- [10] Psi, "Psi," <https://thepsiapp.com/>, 2022, [Online; accessed 27-Jan-2022].
- [11] TOBU, "Tobu," www.tobu.life, 2022, [Online; accessed 27-Jan-2022].
- [12] JaseciLabs, "Jaseci documentation," <https://docs.jaseci.org/>, 2023, [Online; accessed 03-May-2023].
- [13] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "Tinybert: Distilling BERT for natural language understanding," *CoRR*, vol. abs/1909.10351, 2019. [Online]. Available: <http://arxiv.org/abs/1909.10351>

- [14] I. Casanueva, T. Temcinas, D. Gerz, M. Henderson, and I. Vulic, "Efficient intent detection with dual sentence encoders," *CoRR*, vol. abs/2003.04807, 2020. [Online]. Available: <https://arxiv.org/abs/2003.04807>
- [15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [16] D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y. Sung, B. Strope, and R. Kurzweil, "Universal sentence encoder," *CoRR*, vol. abs/1803.11175, 2018. [Online]. Available: <http://arxiv.org/abs/1803.11175>
- [17] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.