Surveyed Papers Table

List of Papers

Below is a table listing the referenced papers in this study.

Citation Key	Reference
fan2023large	[Fan et al.(2023)]
cheng2023ai	[Cheng et al.(2023)]
josefsson2017root	[Josefsson(2017)]
saha2022mining	[Saha and Hoi(2022)]
masood2019aiops	[Masood et al.(2019)]
zhao2023knn	[Zhao et al.(2023a)]
thirunavukarasu2023large	[Thirunavukarasu et al.(2023)]
clusmann2023future	[Clusmann et al.(2023)]
soldani2022anomaly	[Soldani and Brogi(2022)]
wang2024comprehensive	[Wang and Qi(2024)]
zhang2024failure	[Zhang et al.(2024d)]
remil2024aiops	[Remil et al.(2024)]
su2024large	[Su et al.(2024)]
urlana2024llms	[Urlana et al.(2024)]
he2024intelligent	[He et al.(2024)]
kang2023preliminary	[Kang et al.(2023)]
zhang2024automated	[Zhang et al.(2024a)]
chen2023automatic	[Chen et al.(2023)]
zhang2023pace	[Zhang et al.(2023)]
goel2024x	[Goel et al.(2024)]
huang2024faultprofit	[Huang et al.(2024)]
ahmed2023recommending	[Ahmed et al.(2023)]
roy2024exploring	[Roy et al.(2024)]
sarda2024augmenting	[Sarda et al.(2024b)]
wang2023rcagent	[Wang et al.(2023a)]
zhang2024mabc	[Zhang et al.(2024b)]
li2025coca	[Li et al.(2025)]
han2024potential	[Han et al.(2024)]
toro2024retrieval	[Toro Isaza et al.(2024)]
chen2025aiopslab	[Chen et al.(2025)]
zhao2023survey	[Zhao et al.(2023b)]
chang2023survey	[Chang et al.(2023)]
vitui2025empowering	[Vitui and Chen(2025)]
jiang2023xpert	[Jiang et al.(2023)]
an2024nissist	[An et al.(2024)]
somashekar2023enhancing	[Somashekar and Kumar(2023)]
hamadanian2023holistic	[Hamadanian et al.(2023)]
las2024llexus	[Las-Casas et al.(2024)]
samanta2023insightssumm	[Samanta et al.(2023)]

Continued on next page

Table 1 – continued from previous page

Citation Key	Reference
geng2024large	[Geng et al.(2024)]
lanciano2023analyzing	[Lanciano et al.(2023)]
hu2024inference	[Hu et al.(2024)]
wang2024towards	[Wang et al.(2024)]
he2023unicron	[He et al.(2023)]
sarda2023adarma	[Sarda et al.(2023)]
baghdasaryan4690081knowledge	[Baghdasaryan et al.([n. d.])]
nijkamp2022codegen	[Nijkamp et al.(2022)]
touvron2023llama	[Touvron et al.(2023)]
ouyang2022training	[Ouyang et al.(2022)]
manna1971toward	[Manna and Waldinger(1971)]
gulwani2017program	[Gulwani et al.(2017)]
boateng2024survey	[Boateng et al.(2024)]
petke2017genetic	[Petke et al.(2017)]
goues2019automated	[Goues et al.(2019)]
khlaisamniang2023generative	[Khlaisamniang et al.(2023)]
wang2021codet5	[Wang et al.(2021)]
xu2023cloudeval	[Xu et al.(2023)]
chanus2023llm	[Chanus and Aubertin(2023)]
sarda2023leveraging	[Sarda(2023)]
hadi2023survey	[Hadi et al.(2023)]
kratzkedon	[Kratzke and Drews([n. d.])]
pujar2023automated	[Pujar et al.(2023)]
komal2023adarma	[Komal et al.(2023)]
chen2021evaluating	[Chen et al.(2021)]
namrud2024kubeplaybook	[Namrud et al.(2024)]
toprani2025llm	[Toprani and Madisetti(2025)]
sarda2024leveraging	[Sarda et al.(2024a)]
ahmed2024automatic	[Ahmed et al.(2024)]
zhang2024survey	[Zhang et al.(2024c)]
shi2023shellgpt	[Shi et al.(2023)]
xue2023db	[Xue et al.(2023)]
wang2023low	[Wang et al.(2023b)]
cao2024managing	[Cao et al.(2024)]

References

[Ahmed et al.(2023)] Toufique Ahmed, Supriyo Ghosh, Chetan Bansal, Thomas Zimmermann, Xuchao Zhang, and Saravan Rajmohan. 2023. Recommending root-cause and mitigation steps for cloud incidents using large language models. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 1737–1749.

[Ahmed et al.(2024)] Toufique Ahmed, Kunal Suresh Pai, Premkumar Devanbu, and Earl Barr. 2024. Automatic semantic augmentation of language model prompts (for code summarization). In *Proceedings of the IEEE/ACM 46th international conference on software engineering*. 1–13.

[An et al.(2024)] Kaikai An, Fangkai Yang, Liqun Li, Zhixing Ren, Hao Huang, Lu Wang, Pu Zhao, Yu Kang, Hua Ding, Qingwei Lin, et al. 2024. Nissist: An Incident Mitigation Copilot based on Troubleshooting Guides. *arXiv preprint arXiv:2402.17531* (2024).

[Baghdasaryan et al.([n. d.])] Ashot Baghdasaryan, Tigran Bunarjyan, Arnak Poghosyan, Ashot Harutyunyan, and Jad El-Zein. [n. d.]. Knowledge Retrieval and Diagnostics in Cloud Services with Large Language Models. *Available at SSRN 4690081* ([n. d.]).

[Boateng et al.(2024)] Gordon Owusu Boateng, Hani Sami, Ahmed Alagha, Hanae Elmekki, Ahmad Hammoud, Rabeb Mizouni, Azzam Mourad, Hadi Otrok, Jamal Bentahar, Sami Muhaidat, et al. 2024. A Survey on Large Language Models for Communication, Network, and Service Management: Application Insights, Challenges, and Future Directions. *arXiv preprint arXiv:2412.19823* (2024).

[Cao et al.(2024)] Charles Cao, Feiyi Wang, Lisa Lindley, and Zejiang Wang. 2024. Managing Linux servers with LLM-based AI agents: An empirical evaluation with GPT4. *Machine Learning with Applications* 17 (2024), 100570.

[Chang et al.(2023)] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology* (2023).

[Chanus and Aubertin(2023)] Thibault Chanus and Michael Aubertin. 2023. LLM and Infrastructure as a Code use case. *arXiv preprint arXiv:2309.01456* (2023).

[Chen et al.(2021)] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).

[Chen et al.(2025)] Yinfang Chen, Manish Shetty, Gagan Somashekar, Minghua Ma, Yogesh Simmhan, Jonathan Mace, Chetan Bansal, Rujia Wang, and Saravan Rajmohan. 2025. AIOpsLab: A Holistic Framework to Evaluate AI Agents for Enabling Autonomous Clouds. *arXiv preprint arXiv:2501.06706* (2025).

[Chen et al.(2023)] Yinfang Chen, Huaibing Xie, Minghua Ma, Yu Kang, Xin Gao, Liu Shi, Yunjie Cao, Xuedong Gao, Hao Fan, Ming Wen, et al. 2023. Automatic Root Cause Analysis via Large Language Models for Cloud Incidents. *arXiv preprint arXiv:2305.15778* (2023).

[Cheng et al.(2023)] Qian Cheng, Doyen Sahoo, Amrita Saha, Wenzhuo Yang, Chenghao Liu, Gerald Woo, Manpreet Singh, Silvio Saverese, and Steven CH Hoi. 2023. Ai for it operations (aiops) on cloud platforms: Reviews, opportunities and challenges. *arXiv preprint arXiv:2304.04661* (2023).

[Clusmann et al.(2023)] Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, et al. 2023. The future landscape of large language models in medicine. *Communications medicine* 3, 1 (2023), 141.

[Fan et al.(2023)] Angela Fan, Beliz Gokkaya, Mark Harman, Mitya Lyubarskiy, Shubho Sengupta, Shin Yoo, and Jie M Zhang. 2023. Large language models for software engineering: Survey and open problems. *arXiv preprint arXiv:2310.03533* (2023).

[Geng et al.(2024)] Mingyang Geng, Shangwen Wang, Dezun Dong, Haotian Wang, Ge Li, Zhi Jin, Xiaoguang Mao, and Xiangke Liao. 2024. Large language models are few-shot summarizers: Multi-intent comment generation via in-context learning. In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering*. 1–13.

[Goel et al.(2024)] Drishti Goel, Fiza Husain, Aditya Singh, Supriyo Ghosh, Anjaly Parayil, Chetan Bansal, Xuchao Zhang, and Saravan Rajmohan. 2024. X-lifecycle learning for cloud incident management using llms. In *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*. 417–428.

[Goues et al.(2019)] Claire Le Goues, Michael Pradel, and Abhik Roychoudhury. 2019. Automated program repair. *Commun. ACM* 62, 12 (2019), 56–65.

[Gulwani et al.(2017)] Sumit Gulwani, Oleksandr Polozov, Rishabh Singh, et al. 2017. Program synthesis. *Foundations and Trends® in Programming Languages* 4, 1-2 (2017), 1–119.

[Hadi et al.(2023)] Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints* (2023).

[Hamadanian et al.(2023)] Pouya Hamadanian, Behnaz Arzani, Sadjad Fouladi, Siva Kesava Reddy Kakarla, Rodrigo Fonseca, Denizcan Billor, Ahmad Cheema, Edet Nkposong, and Ranveer Chandra. 2023. A Holistic View of AI-driven Network Incident Management. In *Proceedings of the 22nd ACM Workshop on Hot Topics in Networks*. 180–188.

[Han et al.(2024)] Yongqi Han, Qingfeng Du, Ying Huang, Jiaqi Wu, Fulong Tian, and Cheng He. 2024. The Potential of One-Shot Failure Root Cause Analysis: Collaboration of the Large Language Model and Small Classifier. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*. 931–943.

[He et al.(2023)] Tao He, Xue Li, Zhibin Wang, Kun Qian, Jingbo Xu, Wenyuan Yu, and Jingren Zhou. 2023. Unicron: Economizing self-healing llm training at scale. *arXiv preprint arXiv:2401.00134* (2023).

[He et al.(2024)] Yuhang He, Yiming Pan, Yong Wang, Shuqian Du, and Qi Xin. 2024. Intelligent Fault Analysis With AIOps Technology. *Journal of Theory and Practice of Engineering Science* 4, 01 (2024), 94–100.

[Hu et al.(2024)] Cunchen Hu, Heyang Huang, Liangliang Xu, Xusheng Chen, Jiang Xu, Shuang Chen, Hao Feng, Chenxi Wang, Sa Wang, Yungang Bao, et al. 2024. Inference without Interference: Disaggregate LLM Inference for Mixed Downstream Workloads. *arXiv preprint arXiv:2401.11181* (2024).

[Huang et al.(2024)] Junjie Huang, Jinyang Liu, Zhuangbin Chen, Zhihan Jiang, Yichen Li, Jiazhen Gu, Cong Feng, Zengyin Yang, Yongqiang Yang, and Michael R Lyu. 2024. Faultprofit: Hierarchical fault profiling of incident tickets in large-scale cloud systems. In *Proceedings of the 46th International Conference on Software Engineering: Software Engineering in Practice*. 392–404.

[Jiang et al.(2023)] Yuxuan Jiang, Chaoyun Zhang, Shilin He, Zhihao Yang, Minghua Ma, Si Qin, Yu Kang, Yingnong Dang, Saravan Rajmohan, Qingwei Lin, et al. 2023. Xpert: Empowering incident management with query recommendations via large language models. *arXiv preprint arXiv:2312.11988* (2023).

[Josefsson(2017)] Tim Josefsson. 2017. Root-cause analysis throughmachine learning in the cloud.

[Kang et al.(2023)] Sungmin Kang, Gabin An, and Shin Yoo. 2023. A preliminary evaluation of llm-based fault localization. *arXiv preprint arXiv:2308.05487* (2023).

[Khlaisamniang et al.(2023)] Pitikorn Khlaisamniang, Prachaya Khomduean, Kriangkrai Saetan, and Supasin Wonglapsuwan. 2023. Generative AI for self-healing systems. In *2023 18th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*. IEEE, 1–6.

[Komal et al.(2023)] Sarda Komal, Namrud Zakeya, Rouf Raphael, Ahuja Harit, Rasolroveicy Mohammadreza, Litoiu Marin, Shwartz Larisa, and Watts Ian. 2023. ADARMA auto-detection and auto-remediation of microservice anomalies by leveraging large language models. In *Proceedings of the 33rd Annual International Conference on Computer Science and Software Engineering*. 200–205.

[Kratzke and Drews([n. d.])] Nane Kratzke and André Drews. [n. d.]. Don't train, just prompt: Towards a Prompt Engineering Approach for a More Generative Container Orchestration Management. ([n. d.]).

[Lanciano et al.(2023)] Giacomo Lanciano, Manuel Stein, Volker Hilt, Tommaso Cucinotta, et al. 2023. Analyzing declarative deployment code with large language models. *CLOSER* 2023 (2023), 289–296.

[Las-Casas et al.(2024)] Pedro Las-Casas, Alok Gautum Kumbhare, Rodrigo Fonseca, and Sharad Agarwal. 2024. LLexus: an AI agent system for incident management. *ACM SIGOPS Operating Systems Review* 58, 1 (2024), 23–36.

[Li et al.(2025)] Yichen Li, Yulun Wu, Jinyang Liu, Zhihan Jiang, Zhuangbin Chen, Guangba Yu, and Michael R Lyu. 2025. COCA: Generative Root Cause Analysis for Distributed Systems with Code Knowledge. *arXiv preprint arXiv:2503.23051* (2025).

[Manna and Waldinger(1971)] Zohar Manna and Richard J Waldinger. 1971. Toward automatic program synthesis. *Commun. ACM* 14, 3 (1971), 151–165.

[Masood et al.(2019)] Adnan Masood, Adnan Hashmi, Adnan Masood, and Adnan Hashmi. 2019. AIOps: predictive analytics & machine learning in operations. *Cognitive computing recipes: Artificial intelligence solutions using microsoft cognitive services and TensorFlow* (2019), 359–382.

[Namrud et al.(2024)] Zakeya Namrud, Komal Sarda, Marin Litoiu, Larisa Shwartz, and Ian Watts. 2024. Kubeplaybook: A repository of ansible playbooks for kubernetes auto-remediation with llms. In *Companion of the 15th ACM/SPEC International Conference on Performance Engineering*. 57–61.

[Nijkamp et al.(2022)] Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. 2022. Codegen: An open large language model for code with multi-turn program synthesis. *arXiv preprint arXiv:2203.13474* (2022).

[Ouyang et al.(2022)] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[Petke et al.(2017)] Justyna Petke, Saemundur O Haraldsson, Mark Harman, William B Langdon, David R White, and John R Woodward. 2017. Genetic improvement of software: a comprehensive survey. *IEEE Transactions on Evolutionary Computation* 22, 3 (2017), 415–432.

[Pujar et al.(2023)] Saurabh Pujar, Luca Buratti, Xiaojie Guo, Nicolas Dupuis, Burn Lewis, Sahil Suneja, Atin Sood, Ganesh Nalawade, Matt Jones, Alessandro Morari, et al. 2023. Automated code generation for information technology tasks in yaml through large language models. In *2023 60th ACM/IEEE Design Automation Conference (DAC)*. IEEE, 1–4.

[Remil et al.(2024)] Youcef Remil, Anes Bendimerad, Romain Mathonat, and Mehdi Kaytoue. 2024. Aiops solutions for incident management: Technical guidelines and a comprehensive literature review. *arXiv preprint arXiv:2404.01363* (2024).

[Roy et al.(2024)] Devjeet Roy, Xuchao Zhang, Rashi Bhave, Chetan Bansal, Pedro Las-Casas, Rodrigo Fonseca, and Saravan Rajmohan. 2024. Exploring llm-based agents for root cause analysis. In *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*. 208–219.

[Saha and Hoi(2022)] Amrita Saha and Steven CH Hoi. 2022. Mining root cause knowledge from cloud service incident investigations for aiops. In *Proceedings of the 44th International Conference on Software Engineering: Software Engineering in Practice*. 197–206.

[Samanta et al.(2023)] Suranjana Samanta, Oishik Chatterjee, Neil Boyette, Guangya Liu, and Prateeti Mohapatra. 2023. InsightsSumm-Summarization of ITOps Incidents Through In-Context Prompt Engineering. In *2023 IEEE 16th International Conference on Cloud Computing (CLOUD)*. IEEE, 290–292.

[Sarda(2023)] Komal Sarda. 2023. Leveraging Large Language Models for Auto-remediation in Microservices Architecture. In *2023 IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C)*. IEEE, 16–18.

[Sarda et al.(2024a)] Komal Sarda, Zakeya Namrud, Marin Litoiu, Larisa Shwartz, and Ian Watts. 2024a. Leveraging Large Language Models for the Auto-remediation of Microservice Applications: An Experimental Study. In *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*. 358–369.

[Sarda et al.(2023)] Komal Sarda, Zakeya Namrud, Raphael Rouf, Harit Ahuja, Mohammadreza Rasolroveicy, Marin Litoiu, Larisa Shwartz, and Ian Watts. 2023. Adarma auto-detection and auto-remediation of microservice anomalies by leveraging large language models. In *Proceedings of the 33rd Annual International Conference on Computer Science and Software Engineering*. 200–205.

[Sarda et al.(2024b)] Komal Sarda, Zakeya Namrud, Ian Watts, Larisa Shwartz, Seema Nagar, Prateeti Mohapatra, and Marin Litoiu. 2024b. Augmenting Automatic Root-Cause Identification with Incident Alerts Using LLM. In *2024 34th International Conference on Collaborative Advances in Software and COmputiNg (CASCON)*. IEEE, 1–10.

[Shi et al.(2023)] Jie Shi, Sihang Jiang, Bo Xu, Jiaqing Liang, Yanghua Xiao, and Wei Wang. 2023. Shellgpt: Generative pre-trained transformer model for shell language understanding. In *2023 IEEE 34th International Symposium on Software Reliability Engineering (ISSRE)*. IEEE, 671–682.

[Soldani and Brogi(2022)] Jacopo Soldani and Antonio Brogi. 2022. Anomaly detection and failure root cause analysis in (micro) service-based cloud applications: A survey. *ACM Computing Surveys (CSUR)* 55, 3 (2022), 1–39.

[Somashekar and Kumar(2023)] Gagan Somashekar and Rajat Kumar. 2023. Enhancing the configuration tuning pipeline of large-scale distributed applications using large language models (idea paper). In *Companion of the 2023 ACM/SPEC International Conference on Performance Engineering*. 39–44.

[Su et al.(2024)] Jing Su, Chufeng Jiang, Xin Jin, Yuxin Qiao, Tingsong Xiao, Hongda Ma, Rong Wei, Zhi Jing, Jiajun Xu, and Junhong Lin. 2024. Large Language Models for Forecasting and Anomaly Detection: A Systematic Literature Review. *arXiv preprint arXiv:2402.10350* (2024).

[Thirunavukarasu et al.(2023)] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine* 29, 8 (2023), 1930–1940.

[Toprani and Madisetti(2025)] Dheer Toprani and Vijay K Madisetti. 2025. LLM Agentic Workflow for Automated Vulnerability Detection and Remediation in Infrastructure-as-Code. *IEEE Access* (2025).

[Toro Isaza et al.(2024)] Paulina Toro Isaza, Michael Nidd, Noah Zheutlin, Jae-wook Ahn, Chidansh Amitkumar Bhatt, Yu Deng, Ruchi Mahindru, Martin Franz, Hans Florian, and Salim Roukos. 2024. Retrieval Augmented Generation-Based Incident Resolution Recommendation System for IT Support. *arXiv e-prints* (2024), arXiv–2409.

[Touvron et al.(2023)] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[Urlana et al.(2024)] Ashok Urlana, Charaka Vinayak Kumar, Ajeet Kumar Singh, Bala Mallikarjunarao Garlapati, Srinivasa Rao Chalamala, and Rahul Mishra. 2024. LLMs with Industrial Lens: Deciphering the Challenges and Prospects–A Survey. *arXiv preprint arXiv:2402.14558* (2024).

[Vitui and Chen(2025)] Arthur Vitui and Tse-Hsun Chen. 2025. Empowering AIOps: Leveraging Large Language Models for IT Operations Management. *arXiv preprint arXiv:2501.12461* (2025).

[Wang et al.(2023b)] Sheng-Kai Wang, Shang-Pin Ma, Chen-Hao Chao, and Guan-Hong Lai. 2023b. Low-code ChatOps for microservices systems using service composition. In *2023 IEEE International Conference on e-Business Engineering (ICEBE)*. IEEE, 55–62.

[Wang and Qi(2024)] Tingting Wang and Guilin Qi. 2024. A Comprehensive Survey on Root Cause Analysis in (Micro) Services: Methodologies, Challenges, and Trends. *arXiv preprint arXiv:2408.00803* (2024).

[Wang et al.(2024)] Yuxin Wang, Yuhan Chen, Zeyu Li, Zhenheng Tang, Rui Guo, Xin Wang, Qiang Wang, Amelie Chi Zhou, and Xiaowen Chu. 2024. Towards Efficient and Reliable LLM Serving: A Real-World Workload Study. *arXiv preprint arXiv:2401.17644* (2024).

[Wang et al.(2021)] Yue Wang, Weishi Wang, Shafiq Joty, and Steven CH Hoi. 2021. Codet5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. *arXiv preprint arXiv:2109.00859* (2021).

[Wang et al.(2023a)] Zefan Wang, Zichuan Liu, Yingying Zhang, Aoxiao Zhong, Lunting Fan, Lingfei Wu, and Qingsong Wen. 2023a. RCAgent: Cloud Root Cause Analysis by Autonomous Agents with Tool-Augmented Large Language Models. *arXiv preprint arXiv:2310.16340* (2023).

[Xu et al.(2023)] Yifei Xu, Yuning Chen, Xumiao Zhang, Xianshang Lin, Pan Hu, Yunfei Ma, Songwu Lu, Wan Du, Z Morley Mao, Ennan Zhai, et al. 2023. CloudEval-YAML: A Realistic and Scalable Benchmark for Cloud Configuration Generation. (2023).

[Xue et al.(2023)] Siqiao Xue, Caigao Jiang, Wenhui Shi, Fangyin Cheng, Keting Chen, Hongjun Yang, Zhiping Zhang, Jianshan He, Hongyang Zhang, Ganglin Wei, et al. 2023. Db-gpt: Empowering database interactions with private large language models. *arXiv preprint arXiv:2312.17449* (2023).

[Zhang et al.(2023)] Dylan Zhang, Xuchao Zhang, Chetan Bansal, Pedro Las-Casas, Rodrigo Fonseca, and Saravan Rajmohan. 2023. PACE: Prompting and Augmentation for Calibrated Confidence Estimation with GPT-4 in Cloud Incident Root Cause Analysis. *arXiv preprint arXiv:2309.05833* (2023).

[Zhang et al.(2024c)] Lingzhe Zhang, Tong Jia, Mengxi Jia, Yifan Wu, Aiwei Liu, Yong Yang, Zhonghai Wu, Xuming Hu, Philip S Yu, and Ying Li. 2024c. A survey of aiops for failure management in the era of large language models. *arXiv preprint arXiv:2406.11213* (2024).

[Zhang et al.(2024d)] Shenglin Zhang, Sibo Xia, Wenzhao Fan, Binpeng Shi, Xiao Xiong, Zhenyu Zhong, Minghua Ma, Yongqian Sun, and Dan Pei. 2024d. Failure diagnosis in microservice systems: A comprehensive survey and analysis. *ACM Transactions on Software Engineering and Methodology* (2024).

[Zhang et al.(2024b)] Wei Zhang, Hongcheng Guo, Jian Yang, Yi Zhang, Chaoran Yan, Zhoujin Tian, Hangyuan Ji, Zhoujun Li, Tongliang Li, Tieqiao Zheng, et al. 2024b. mABC: multi-Agent Blockchain-Inspired Collaboration for root cause analysis in micro-services architecture. *arXiv preprint arXiv:2404.12135* (2024).

[Zhang et al.(2024a)] Xuchao Zhang, Supriyo Ghosh, Chetan Bansal, Rujia Wang, Minghua Ma, Yu Kang, and Saravan Rajmohan. 2024a. Automated Root Causing of Cloud Incidents using In-Context Learning with GPT-4. *arXiv preprint arXiv:2401.13810* (2024).

[Zhao et al.(2023a)] Wenting Zhao, Ye Liu, Yao Wan, Yibo Wang, Qingyang Wu, Zhongfen Deng, Jiangshu Du, Shuaiqi Liu, Yunlong Xu, and Philip S Yu. 2023a. Knn-icl: Compositional task-oriented parsing generalization with nearest neighbor in-context learning. *arXiv preprint arXiv:2312.10771* (2023).

[Zhao et al.(2023b)] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023b. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).