
White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Leeds.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/246584/

Version: Accepted Version

Proceedings Paper:
Fan, Y., Zaidi, S.A.R. and Hafeez, M. (Accepted: 2026) Telecom-Specific Small Language Models for Root-Cause Analysis. In: Proceedings of 2026 IEEE Globecom. 2026 IEEE Globecom, 07-11 Dec 2026, Macau S.A.R., China. IEEE. (In Press)

This is an author produced version of a conference paper accepted for publication in Proceedings of IEEE GLOBECOM 2026, made available via the University of Leeds Research Outputs Policy under the terms of the Creative Commons Attribution License (CC-BY), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.





Telecom-Specific Small Language Models for Root-Cause Analysis

Yejing Fan, Syed Ali Raza Zaidi, Maryam Hafeez
School of Electronic and Electrical Engineering
University of Leeds
Leeds, United Kingdom
Email: {Y.Fan1, S.A.Zaidi, M.Hafeez}@leeds.ac.uk

***Abstract*—Fault localization is a fundamental task for reliable and automated telecommunication network operation, while root-cause analysis (RCA) goes a step further by identifying the underlying cause of an observed network fault. Recent large language model (LLM)-based RCA approaches provide strong diagnostic capability but can be computationally expensive for deployment. This paper investigates whether such capability can be distilled into a compact small language model (SLM). We formulate TeleLogs RCA as an eight-class closed-set classification problem and distil diagnostic knowledge from an OTel-LLM-8.3B teacher into a LoRA-fine-tuned Qwen2.5-1.5B student with a sequence-classification head. Under a strict train/validation/test protocol, the two-stage knowledge-distillation (KD) student achieves 84.72% accuracy and 83.59% macro-F1. Its dominant failure is confusion between excessive serving-cell downtilt (C1) and higher-throughput neighbouring-cell coverage (C3), producing 97 direct class swaps. We therefore introduce a lightweight evidence-aware specialist that explicitly binds drive-test measurements to engineering parameters. The hierarchical system reaches 94.44% accuracy using four interpretable features and 95.49% using the full structured representation, reducing C1/C3 swaps from 97 to 4. Controlled perturbations further reveal severe SLM sensitivity to telemetry serialization, while the structured evidence remains invariant for all 864 test samples, localizing the remaining robustness bottleneck to SLM-based routing. The student also reduces peak GPU memory by 4.53× and improves throughput by 1.51× over the teacher. These results demonstrate both the deployment potential of compact telecom-specific SLMs and the importance of explicit relational evidence processing for reliable RCA.**

***Index Terms*—telecommunication root-cause analysis, small language model, knowledge distillation, LoRA, TeleLogs, model compression, network automation.**

I. Introduction

Modern 5G and emerging 6G networks are increasingly software-defined, telemetry-rich, and operationally heterogeneous. Dense radio deployments, mobility, inter-cell interference, handover control, and dynamic resource allocation create fault signatures that may appear similar at the key performance indicator level [1]–[3]. Therefore, detecting a degradation in throughput or coverage is only the first step: network operation and management systems must also infer the most plausible underlying cause. Root-cause analysis (RCA) is therefore a central capability for self-optimizing and self-healing mobile networks. Conventional RCA solutions rely on expert rules, fault trees, statistical learning, or graph-based dependency modelling [4]–[6]. Although effective in well-defined settings, these approaches can be costly to maintain as network configurations, fault types, and operational evidence evolve.

Language models provide a complementary route because they can jointly process textual instructions, structured measurements, and domain descriptions. Recent studies have explored LLM-based agents for incident diagnosis, tool-assisted cloud RCA, and procedure-guided multi-agent troubleshooting [7]–[10]. In the mobile-network domain, Sana *et al.* introduced TeleLogs and showed that generic instruction-tuned and reasoning models perform poorly on specialised 5G troubleshooting unless they are adapted to the task [11]. Their results demonstrate the value of domain-specific post-training, but also expose a deployment challenge: high-capacity reasoning models require substantial memory and inference resources.

This challenge is particularly relevant to operational RCA, where the same bounded diagnostic decisions are executed repeatedly and may need to run close to network-management platforms. Recent surveys and model studies show that small language models (SLMs) can offer lower latency, smaller memory footprints, and easier domain adaptation than general-purpose LLMs [12]–[14]. Belcak *et al.* further argue that many agentic workloads are narrow, repetitive, and therefore better matched to specialised SLMs than to monolithic generalist models [15]. For telecom RCA, this suggests that the operational decision module does not necessarily need open-domain conversational breadth; instead, it needs reliable discrimination among a fixed set of technically similar fault causes.

Knowledge distillation (KD) provides a natural mechanism for transferring task knowledge from a high-capacity teacher to a smaller deployable student [16]. In the present work, the teacher remains the stronger telecom RCA classifier, while the central question is whether its fine-grained class-boundary information can be retained by a compact student through soft-label supervision and parameter-efficient adaptation [17].

We investigate not only how much telecom RCA capability survives compression, but also what reasoning capability is lost. An OTel-LLM-8.3B classifier is distilled into a Qwen2.5-1.5B sequence classifier using LoRA. Error analysis shows that the compact student struggles primarily with cross-table discrimination between coverage- and neighbour-related causes. We therefore augment the SLM with a lightweight structured-evidence specialist and evaluate the resulting hierarchy under both the standard TeleLogs test set and controlled

evidence-order perturbations. The main contributions are:

- a strict two-stage KD evaluation of a 1.56B telecom RCA classifier, achieving 84.72% accuracy on the 864-sample TeleLogs test set;
- an evidence-aware C1/C3 specialist that raises full eight-class accuracy to 94.44% with four interpretable features and 95.49% with the complete feature set; and
- a controlled robustness study showing that structured telecom evidence is permutation-invariant while the compact SLM remains highly sensitive to telemetry serialization, identifying SLM-based routing as the principal remaining robustness bottleneck.

II. Related Work

A. Root-Cause Analysis in Communication Networks

Traditional network RCA combines domain knowledge with statistical or graph-based inference. Wireless-network studies have applied fault localisation, graph neural networks, and spatial dependency modelling to KPI-driven diagnosis [4]–[6], [18]. These approaches provide structured inference but are generally designed around predefined features or graph structures.

B. Language Models for RCA and Operational Diagnosis

LLM-based systems extend RCA from fixed feature pipelines toward natural-language reasoning, tool use, and multi-agent workflows [7]–[9]. More generally, agent architectures decompose complex objectives into planning, execution, tool calling, and reflection stages [19], [20]. These systems improve flexibility and human-readable interaction, but their reliance on large generative models may be unnecessary when the downstream decision is a bounded classification problem. The closest work is the TeleLogs study by Sana *et al.* [11]. It formulates 5G troubleshooting as the identification and explanation of one of eight root causes and proposes supervised fine-tuning followed by reinforcement learning for reasoning models. Their domain-adapted Qwen2.5-RCA models substantially outperform generic reasoning baselines, establishing TeleLogs as a challenging specialised benchmark. Our work addresses a different operational objective: rather than generating long reasoning traces with models up to 32B parameters, we compress the final C1-C8 diagnostic capability into a 1.5B sequence classifier and evaluate both predictive quality and inference cost.

C. Small Language Models and Knowledge Distillation

SLMs have gained attention for on-device inference, privacy, lower serving cost, and rapid task specialisation [12]–[14]. From a systems perspective, Belcak *et al.* argue for SLM-first agentic designs in which specialised models handle narrow recurring calls and larger models are reserved for genuinely open-ended tasks [15]. Telecom RCA fits this pattern because the evidence format is structured and the final action space is fixed. KD transfers information from a teacher distribution to a smaller student, allowing the student to learn inter-class similarity that is absent from one-hot targets [16]. Parameter-efficient methods such as LoRA further reduce adaptation cost by updating low-rank matrices rather than all backbone parameters [17]. Our focus differs from general model compression by examining whether fine-grained cross-table telecom reasoning survives distillation and how explicit structured evidence can compensate when it does not.

III. Methodology

A. Task and Compact Student

TeleLogs contains 2400 training and 864 official test samples, each combining eight candidate RCA definitions with 5G drive-test measurements and engineering parameters [11]. The eight classes describe downtilt, overshooting, neighbouring-cell throughput, overlapping coverage, frequent handovers, PCI-mod-30 interference, excessive vehicle speed, and insufficient scheduled resource blocks. We formulate the task as direct eight-class classification,

$$\hat{y} = \arg \max_{c \in \{1,\dots,8\}} p_\theta(c|x). \tag{1}$$

The teacher model is an OTel-LLM-8.3B sequence-classification model trained for TeleLogs RCA [21]. The student is based on Qwen2.5-1.5B-Instruct [22] with a sequence-classification head for eight RCA classes. To reduce training cost, we fine-tune the model using LoRA adapters [17] applied to attention and feed-forward projection layers, while the classification head is saved as a trainable module. The final student has 1.56B total parameters, with 18.48M trainable parameters during LoRA fine-tuning.

B. Two-Stage Knowledge Distillation

For input x_i, let q_i^T and q_i^S denote the teacher and student class distributions softened using temperature τ. The student minimizes

$$\mathcal{L} = \alpha \mathcal{L}_{\mathrm{CE}} + (1-\alpha)\tau^2 D_{\mathrm{KL}}\left(q_i^T \| q_i^S\right), \tag{2}$$

where $\alpha = 0.5$ and $\tau = 2$.

Training proceeds in two stages. First, the LoRA-adapted student is trained using hard C1-C8 labels to establish task alignment. The resulting checkpoint then initializes Stage 2, where Eq. (2) uses cached teacher probabilities to refine the decision boundaries.

C. Evidence-Aware Hierarchical RCA

The strict student exhibits concentrated confusion between C1 (excessive serving-cell downtilt) and C3 (higher-throughput neighbouring cell). We therefore extract structured evidence by associating drive-test and engineering records through PCI rather than table-row position. The Core-4 representation contains pre-handover total tilt, mean pre-handover RSRP, mean post-handover throughput, and throughput gain. A second variant uses 23 features covering antenna configuration, radio quality, throughput, neighbour margins, mobility, and resource allocation. A logistic-regression specialist ($C = 0.1$) is trained only on C1/C3 training samples after median imputation and standardization.

The SLM remains the primary eight-class model and activates the specialist only when its predicted class is C1 or C3:

$$\hat{y} = \begin{cases} g_\phi(\mathbf{s}(x)), & \hat{y}_{\text{SLM}} \in \{\text{C1}, \text{C3}\}, \\ \hat{y}_{\text{SLM}}, & \text{otherwise.} \end{cases} \tag{3}$$

IV. Experimental Setup

A. Evaluation Setup

We evaluate on the released TeleLogs train/test split with 2400 training samples and 864 test samples. The original 2400 training examples are divided into 2160 training and 240 validation samples using a stratified split with random seed 42. All hyperparameter decisions and checkpoint selection are performed using the training and validation subsets only. The official 864-sample test set is evaluated only after the model checkpoint has been fixed. For the two-stage KD student, the checkpoint with the highest validation macro-F1 is selected for final evaluation.

The maximum sequence length is set to 5000 tokens. Using the Qwen tokenizer, the test prompts have an average length of 2359 tokens, a median length of 2356 tokens, a 95th-percentile length of 2681 tokens, and a maximum length of 2857 tokens. Consequently, all 864 test prompts are fully contained within the 5000-token inference window and no test sample is truncated.

All experiments were conducted on a single NVIDIA DGX Spark equipped with the GB10 Grace Blackwell Superchip, a Blackwell-architecture GPU, a 20-core Arm CPU, and 128 GB of coherent LPDDR5x unified memory. We report accuracy, macro-F1, confusion matrices, latency, throughput, and peak allocated GPU memory.

B. Controlled Robustness Evaluation

To examine whether the compact RCA model relies on superficial serialization cues, we construct deterministic perturbations of the same 864 TeleLogs test samples using random seed 2026. These perturbations preserve the ground-truth RCA label, all numerical KPI values, cell identifiers, and the temporal order of the user-plane measurements.

Four conditions are considered: (i) *option order*, which permutes the display order of the C1-C8 candidate definitions while preserving each identifier-definition pair; (ii) *table order*, which swaps the complete user-plane and engineering-parameter sections; (iii) *engineering-row order*, which permutes only the cell rows of the engineering table while preserving every row; and (iv) *combined*, which applies all three perturbations. The resulting benchmark is therefore a TeleLogs-based controlled perturbation test rather than a separate official randomized split.

As a sanity check, structured features are extracted independently from every original and perturbed prompt. For all four perturbations, both the Core-4 and complete 23-feature representations are identical for 864/864 samples, with a maximum numerical difference of zero.

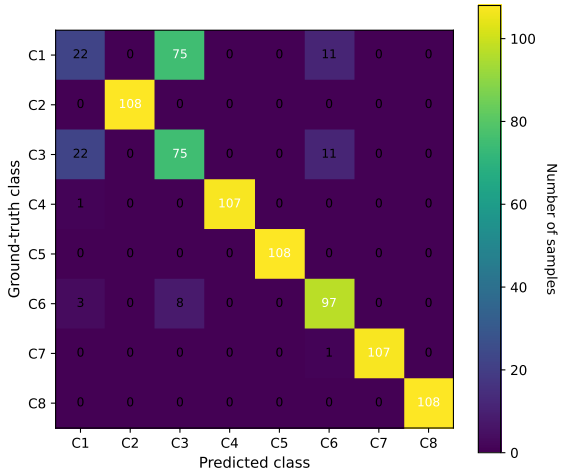


Fig. 1. Confusion matrix of the strictly evaluated two-stage KD student

TABLE I
Strict RCA and hierarchical refinement results.

System	Acc.(%)	F1(%)	C1→C3	C3→C1
Strict two-stage KD	84.72	83.59	75	22
+ Core-4	94.44	94.46	13	0
+ All-feature	**95.49**	**95.55**	**4**	**0**

V. Results and Analysis

A. Strict Two-Stage KD Performance

Under the strict evaluation protocol, the two-stage KD student achieves 84.72% accuracy and 83.59% macro-F1 on the 864-sample official TeleLogs test set, corresponding to 732 correctly classified samples. The corresponding confusion matrix is shown in Fig. 1.

The errors are highly non-uniform across RCA classes. In particular, 75 C1 samples are directly misclassified as C3, while 22 C3 samples are misclassified as C1. Thus, 97 of the student's errors are direct C1/C3 swaps. The student correctly identifies only 22 of 108 C1 samples and 75 of 108 C3 samples, whereas the remaining RCA classes are substantially better separated. This concentration motivates a more detailed examination of the C1/C3 boundary.

B. Evidence Analysis of the C1/C3 Boundary

To determine whether the C1/C3 errors arise from insufficient information in the telemetry, we additionally train binary structured classifiers using only C1/C3 samples. The Core-4 representation reaches 92.59% accuracy and 92.56% macro-F1, while the complete 23-feature representation reaches 98.15% accuracy, 98.15% macro-F1, and 99.49% ROC-AUC. For reference, the eight-class teacher achieves 97.69% accuracy when evaluated on the same C1/C3 subset. Although these binary results are not directly comparable to the eight-class teacher, they indicate that the telemetry contains sufficient discriminative evidence to resolve the C1/C3 boundary.

C. Evidence-Aware Hierarchical RCA

Table I shows that explicit evidence specialization substantially improves the end-to-end eight-class RCA result. The Core-4 specialist increases accuracy from 84.72% to 94.44% and macro-F1 from 83.59% to 94.46%. The all-feature specialist further increases accuracy to 95.49% and macro-F1 to 95.55%.

The improvement is concentrated exactly where expected. The number of direct C1/C3 swaps decreases from 97 for the strict two-stage KD student to 13 with Core-4 and only 4 with the full specialist. C1 correct predictions increase from 22/108 to 84/108 and 93/108, respectively, while C3 correct predictions increase from 75/108 to 97/108 for both hierarchical variants. Overall, the all-feature hierarchy reduces the number of test errors from 132 to 39, corresponding to a 70.5% reduction in classification errors.

The main SLM routes 206 samples to the specialist. Among these, 194 are true C1/C3 samples, while 12 belong to other classes. Twenty-two true C1/C3 samples are not routed because the SLM predicts a different RCA class. For the all-feature specialist, 190 of the 194 routed true C1/C3 samples are classified correctly (97.94%). This indicates that, once the correct boundary is activated, the structured specialist is highly effective; the remaining limitation is primarily the routing decision.

D. Robustness and Failure Localization

Table II reveals a strong distinction between candidate-option ordering and evidence serialization. Reordering the eight RCA candidate definitions has essentially no effect on the main student: accuracy remains exactly 84.72%, while macro-F1 changes only from 83.59% to 83.41%. The hierarchical systems are similarly stable, with the all-feature model decreasing by only 0.12 percentage points in accuracy.

In contrast, changing the serialization of the telemetry causes severe degradation. Swapping the user-plane and engineering tables reduces the main-model accuracy from 84.72% to 18.52%, while permuting only the engineering-table rows reduces it to 23.38%. The combined perturbation results in 18.87% accuracy. These changes cannot be attributed to context truncation: all original and perturbed prompts remain below 2857 tokens and are therefore fully contained within the 5000-token inference window.

The degradation cannot also be attributed to altered physical evidence. For each of the 864 samples, both the Core-4 and complete 23-feature representations extracted from every perturbed prompt are numerically identical to those obtained from the original prompt, with a maximum absolute feature difference of zero. Therefore, the structured evidence representation is invariant to the tested permutations.

The routing statistics further localize the failure. Under the original input order, 194 of the 216 true C1/C3 samples are routed to the specialist. Under table-order perturbation, none of the 216 true C1/C3 samples is routed; engineering-row and combined perturbations similarly route zero C1/C3 samples. Consequently, the hierarchical specialist cannot recover the prediction despite receiving an invariant evidence representation. These results dentify the SLM-based routing stage, rather than the structured specialist itself, as the principal robustness bottleneck.

The contrast between stable option-order performance and severe telemetry-order degradation suggests that the compressed SLM does not merely memorize the physical position of the C1-C8 answer candidates. Instead, it appears sensitive to how heterogeneous network evidence is serialized, suggesting reliance on positional or serialization cues when binding drive-test measurements to engineering parameters.

E. Efficiency and Deployment Cost

As shown in Table III, the distilled student reduces the parameter count from 8.31B to 1.56B, corresponding to a $5.33\times$ reduction. It decreases mean end-to-end latency from 608.56 to 404.25 ms. Throughput increases from 1.64 to 2.47 samples/s, representing a $1.51\times$ improvement. Peak allocated memory is reduced from 15.74 to 3.48 GB, corresponding to a $4.52\times$ reduction. The structured specialist is a logistic-regression model operating on at most 23 scalar features and therefore adds negligible model size relative to the 1.56B-parameter SLM.

F. Task-Level Comparison

For task-level context, the current GSMA Open Telco Leaderboard reports a TeleLogs score of 96.3% for OTel-LLM-8.3B-QnA, while the original TeleLogs study reports 95.86% pass@1 for the 32B Qwen2.5-RCA reasoning model [11], [23]. Our 1.56B hierarchical classifier achieves 95.49% on the official 864-sample TeleLogs test set. Since the GSMA leaderboard and our work use different evaluation interfaces and protocols, these values should be interpreted only as task-level context rather than as a direct ranking comparison.

VI. Discussion

The results suggest that output-level KD preserves broad telecom RCA capability more effectively than fine-grained relational reasoning. The dominant C1/C3 errors require the model to associate serving and neighbouring PCIs across heterogeneous telemetry tables and relate them to antenna configuration, radio quality, and handover-dependent throughput. The strong performance of the structured specialist indicates that this information is present in the input, but is not represented reliably by the compressed SLM. The Core-4 result further shows that much of the lost capability can be recovered using only a small number of physically interpretable telecom features rather than a complex auxiliary model.

The perturbation study exposes a different limitation at the system level. Although the structured evidence is exactly invariant to the tested permutations, specialist activation depends entirely on the SLM prediction. Table and engineering-row permutations therefore cause the router to bypass the specialist, even when the underlying evidence remains unchanged. This identifies robust routing, rather than specialist accuracy, as

TABLE II
CONTROLLED ROBUSTNESS EVALUATION.

Condition	Strict KD Acc.	Core-4 Acc.	All-feature Acc.	True C1/C3 routed
Original	84.72	94.44	95.49	194/216
Option order	84.72	94.44	95.37	192/216
Table order	18.52	18.52	18.52	0/216
Engineering-row order	23.38	23.38	23.38	0/216
Combined	18.87	18.87	18.87	0/216

TABLE III
TEACHER-STUDENT EFFICIENCY COMPARISON.

Metric	OTel Teacher	Student
Parameters	8.31B	1.56B
End-to-end latency	608.56 ms	404.25 ms
Throughput	1.64 samples/s	2.47 samples/s
Peak allocated GPU memory	15.74 GB	3.48 GB

the main target for future improvement. Possible directions include evidence-aware gating, canonical telemetry serialization, and permutation-robust representations, which should be developed on training/validation data and evaluated on unseen perturbations.

From an agentic-network-management perspective, the proposed architecture is best viewed as a compact RCA decision module rather than a complete autonomous agent. It can provide a structured diagnosis to higher-level planning and orchestration components while keeping repeated, bounded telecom reasoning computationally lightweight.

VII. CONCLUSION

This paper presented a compact evidence-aware architecture for closed-set telecom RCA. Under a strict evaluation protocol, the 1.56B two-stage KD student achieves 84.72% accuracy on TeleLogs, while the proposed structured specialist raises full eight-class accuracy to 95.49% and reduces direct C1/C3 swaps from 97 to four. Controlled perturbations show that the structured evidence remains permutation-invariant, while the SLM-based router is sensitive to telemetry serialization, revealing a clear remaining robustness challenge. Together with a 5.33× parameter reduction, 4.52× lower peak GPU memory, and 1.51× higher throughput than the teacher, the results support compact evidence-aware SLMs as practical RCA modules for future automated and agentic network management.

ACKNOWLEDGMENT

This research was supported by UK Research and Innovation (UKRI) through the EPSRC: the Technology Missions Fund project CHEDDAR (EP/Y037421/1). This study does not involve human subjects or sensitive data, and raises no ethical or policy concerns.

REFERENCES

[1] R. Zhang, G. Liu, Y. Liu, C. Zhao, J. Wang, Y. Xu, D. Niyato, J. Kang, Y. Li, S. Mao, S. Sun, X. Shen, and D. I. Kim, "Toward edge general intelligence with agentic ai and agentification: Concepts, technologies, and future directions," *IEEE Communications Surveys Tutorials*, vol. 28, pp. 4285–4318, 2026.

[2] A. Mekrache, A. Ksentini, and C. Verikoukis, "Machine learning in fcaps: Toward enhanced beyond 5g network management," *IEEE Communications Surveys Tutorials*, vol. 26, no. 4, pp. 2769–2797, 2024.

[3] Y. Fan, L. Zhang, K. Li, M. Yang, X. Zhang, R. He, and M. Lu, "Attention-bilstm for timely detection and adaptive classification of emi and iemi in 5g-railways wireless communications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 27, no. 7, pp. 8043–8056, 2026.

[4] T. Zhang, Q. Chen, Y. Jiang, D. Miao, F. Yin, T. Quan, Q. Shi, and Z.-Q. Luo, "ICASSP-SPGC 2022: Root cause analysis for wireless network fault localization," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 9301–9305.

[5] C.-C. Yen, W. Sun, H. Purmehdi, W. Park, K. R. Deshmukh, N. Thakrar, O. Nassef, and A. Jacobs, "Graph neural network based root cause analysis using multivariate time-series KPIs for wireless networks," in *Proc. IEEE/IFIP Network Operations and Management Symposium (NOMS)*, 2022, pp. 1–7.

[6] L. Mata, M. Sousa, P. Vieira, M. P. Queluz, and A. Rodrigues, "On the use of spatial graphs for performance degradation root-cause analysis toward self-healing mobile networks," *IEEE Access*, vol. 12, pp. 20 490–20 508, 2024.

[7] D. Roy, X. Zhang, R. Bhave, C. Bansal, P. Las-Casas, R. Fonseca, and S. Rajmohan, "Exploring LLM-based agents for root cause analysis," in *Companion Proc. 32nd ACM Int. Conf. Foundations of Software Engineering*, 2024, pp. 208–219.

[8] Z. Wang, Z. Liu, Y. Zhang, A. Zhong, J. Wang, F. Yin, L. Fan, L. Wu, and Q. Wen, "RCAgent: Cloud root cause analysis by autonomous agents with tool-augmented large language models," in *Proc. 33rd ACM Int. Conf. Information and Knowledge Management*, 2024, pp. 4966–4974.

[9] C. Pei, Z. Wang, F. Liu, Z. Li, Y. Liu, X. He, R. Kang, T. Zhang, J. Chen, J. Li *et al.*, "Flow-of-action: SOP enhanced LLM-based multi-agent system for root cause analysis," in *Companion Proc. ACM Web Conference*, 2025, pp. 422–431.

[10] Y. Fan, Z. Zhou, Z. Qiao, Y. Sun, and L. Zhang, "Efficiency analysis of decentralized autonomous organization (dao) voting mechanisms," in *2024 IEEE Globecom Workshops (GC Wkshps)*, 2024, pp. 1–6.

[11] M. Sana, N. Piovesan, A. De Domenico, Y. Kang, H. Zhang, M. Debbah, and F. Ayed, "Reasoning language models for root cause analysis in 5g wireless networks," *arXiv preprint arXiv:2507.21974*, 2025.

[12] Z. Lu, X. Li, D. Cai, R. Yi, F. Liu, X. Zhang, N. D. Lane, and M. Xu, "Small language models: Survey, measurements, and insights," *arXiv preprint arXiv:2409.15790*, 2024.

[13] F. Wang, Z. Zhang, X. Zhang, Z. Wu, T. Mo, Q. Lu, W. Wang, R. Li, J. Xu, X. Tang, Q. He, Y. Ma, M. Huang, and S. Wang, "A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with LLMs, and trustworthiness," *preprint arXiv:2411.03350*, 2024.

[14] M. Abdin, J. Aneja, H. Awadalla, A. Awadallah, A. A. Awan, N. Bach, A. Bahree, A. Bakhtiari, J. Bao, H. Behl *et al.*, "Phi-3 technical report: A highly capable language model locally on your phone," *arXiv preprint arXiv:2404.14219*, 2024.

[15] P. Belcak, G. Heinrich, S. Diao, Y. Fu, X. Dong, S. Muralidharan, Y. C. Lin, and P. Molchanov, "Small language models are the future of agentic ai," *arXiv preprint arXiv:2506.02153*, 2025.

[16] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[17] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[18] Y. Fan, L. Zhang, and K. Li, "Emi and iemi impacts on the radio communication network of electrified railway systems: A critical review," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 8, pp. 10 409–10 424, 2023.

[19] T. Masterman, S. Besen, M. Sawtell, and A. Chao, "The landscape of emerging AI agent architectures for reasoning, planning, and tool calling: A survey," *arXiv preprint arXiv:2404.11584*, 2024.

[20] Y. Fan, A. Salama, M. M. Hameed Qazzaz, M. Hafeez, S. A. Raza Zaidi, A. Murad, M. Sharara, and V. P. Talari, "Intent-driven agentic ai for safe energy-efficient open ran optimization via a 5g network-in-a-box digital twin," in *2026 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2026, pp. 1–6.

[21] F. Tavakkoli and collaborators, "OTel-LLM-8.3B-Classification," https://huggingface.co/farbodtavakkoli/OTel-LLM-8.3B-Classification, 2025, hugging Face model card.

[22] Qwen Team, "Qwen2.5-1.5B-Instruct," https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct, 2024, hugging Face model card.

[23] GSMA, "GSMA Open Telco Leaderboard," https://huggingface.co/spaces/GSMA/open-telco-leaderboard, 2026, accessed July 2026.