Towards Edge-deployable Telecom Intelligence with Efficient Small Language Models

Anonymous Authors[1]

Abstract

Deployment of large language models in 6G and O-RAN edge environments is constrained by latency, hardware, and privacy requirements that make cloud-hosted inference impractical. While sub-2B small language models (SLMs) are suitable for local deployment, their general-purpose pretraining provides limited telecommunications-domain knowledge, and existing fine-tuning approaches overlook the diversity of knowledge areas and reasoning types the domain encompasses. We propose lightweight task-specific LoRA fine-tuning, training one compact adapter per task category over a shared frozen backbone, and evaluate on the GSMA ot-full benchmark across six task categories using three SLMs. Our results show that SLMs with lightweight LoRA adapters not only narrow the performance gap with larger models but occasionally surpass them, while each adapter introduces only a small fraction of additional parameters relative to the frozen backbone[1]. These findings suggest that lightweight task-specific adaptation offers a practical and efficient path toward edge-native telecommunications intelligence.

1. Introduction

Integration of language model intelligence into next-generation wireless networks has emerged as a promising direction to automate network management, fault diagnosis, and standards-based reasoning (Yin et al., 2026; Zhou et al., 2024). In 6G and O-RAN architectures, language models could support intelligent RAN applications, including log analysis, configuration assistance, and protocol-level question answering (Chen et al., 2026). Deploying large-scale models in these environments remains impractical due to tight control-loop latency requirements, limited edge hardware resources, and privacy constraints that preclude offloading sensitive network telemetry to external cloud endpoints (Qu et al., 2025).

SLMs in the sub-2B parameter range are a natural candidate for edge deployment (Lu et al., 2025), but their general-purpose pretraining leaves them with limited exposure to telecommunications-specific knowledge. Supervised fine-tuning (SFT) is a natural way to adapt these models to the telecommunications domain (Wei et al., 2021; Ouyang et al., 2022). Many existing approaches treat telecommunications as a single monolithic fine-tuning target, overlooking the diversity of knowledge areas and reasoning types it encompasses (Zou et al., 2026; Zhou et al., 2026). The telecommunications domain spans tasks as varied as standards comprehension, fault diagnosis, numerical computation, table understanding, and document classification — each requiring distinct reasoning capabilities. Training a single adapter across all of these simultaneously risks introducing conflicting optimization signals that degrade per-task performance (Mueller et al., 2024; Yu et al., 2020).

In this work, we show that lightweight LoRA fine-tuning can substantially improve SLMs on telecommunications tasks while preserving their deployment-friendly model size. By updating only a small set of adapter parameters, this approach injects telecommunications-domain knowledge into compact models with minimal additional overhead. We evaluate on the GSMA ot-full benchmark (GSMA, 2025) using SLMs from three model families, and compare them against larger zero-shot reference models.

Our contributions are as follows. First, we show that lightweight task-specific LoRA fine-tuning enables SLMs to narrow the performance gap with larger models on telecommunications tasks. Second, we demonstrate that this is achievable with minimal parameter overhead, as each LoRA adapter remains compact relative to the frozen backbone. Third, we provide a systematic evaluation across seen and unseen data splits and an ablation over LoRA rank, offering practical guidance for deploying lightweight adapters in resource-constrained O-RAN environments.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1]The trained LoRA adapters are available on Hugging Face

2. Methodology

2.1. GSMA ot-full Benchmark

We evaluate on the GSMA ot-full benchmark, a unified telecommunications evaluation suite that consolidates eight task categories: TeleQnA (Maatouk et al., 2026), 6G-Bench (Ferrag et al., 2026), 3GPP-TSG (Zou et al., 2025), ORANBench (Gajjar & Shah, 2025a), srsRANBench (Gajjar & Shah, 2025b), TeleLogs (Sana et al., 2025), TeleMath (Colle et al., 2026), and TeleTables (Ezzakri et al., 2025). These categories span diverse knowledge areas and reasoning types, from standards-based multiple-choice QA and document classification to numerical computation, table comprehension, and fault diagnosis. We focus our primary evaluation on six categories that support a unified exact-match protocol: TeleQnA, 6G-Bench, 3GPP-TSG, ORANBench, srsRANBench, and TeleTables. TeleMath and TeleLogs are excluded and left for future work, as they require task-specific scoring and domain-specific evaluation beyond the shared pipeline used in this study. Table 1 summarizes the evaluated subsets. Since the dataset provides no predefined splits, we partition each subset into 60% training, 20% validation, and 20% test using a fixed random seed. Each adapter is trained and selected on its corresponding splits, and we additionally evaluate on both the training data (seen) and test split (unseen) to assess generalization.

Table 1. Overview of evaluated task subsets in GSMA ot-full

Category	Task Type	Total
TeleQnA	Multiple-choice QA	10,000
6G-Bench	Multiple-choice QA	3,720
3GPP_TSG	Document classification	2,000
ORANBench	Multiple-choice QA	1,500
srsRANBench	Multiple-choice QA	1,500
TeleTables	Multiple-choice QA	500

2.2. Lightweight Task-Specific LoRA Fine-Tuning

Instead of training a single adapter across all task categories simultaneously, we train one lightweight LoRA adapter per category over a shared frozen backbone. This avoids conflicting optimization signals that arise when tasks differ in required knowledge and reasoning patterns, allowing each adapter to specialize within its corresponding data distribution. Since only the adapter parameters are updated while the backbone remains frozen, the number of trainable parameters per adapter is a small fraction of the total model size, making each adapter compact and storage-efficient.

Formally, let θ denote the frozen parameters of the base model. For each category $c \in \{1, \cdots, 6\}$, we introduce a LoRA adapter $\Delta\theta_c = BA$, where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times d}$, and $r \ll d$. Rather than updating the full weight matrix, LoRA constrains the weight update to a low-rank subspace, adding only $2dr$ trainable parameters per target module while keeping the backbone entirely frozen. Each adapter is trained independently by minimizing the training loss over its corresponding training split:

$$\mathcal{L}_c = -\sum_{(x,y)\in\mathcal{D}_c^{\text{train}}} \log P_{\theta+\Delta\theta_c}(y \mid x)$$

Since each adapter is trained exclusively on its corresponding category, gradient updates remain focused on a single task distribution. In many O-RAN deployment scenarios, the task category may be inferred from the requesting xApp/rApp context, requiring no additional routing mechanism beyond what is already implicit in the deployment context. For more ambiguous requests where the category cannot be inferred, a lightweight routing component could be incorporated in future work.

3. Evaluation Setup

3.1. Backbone Models

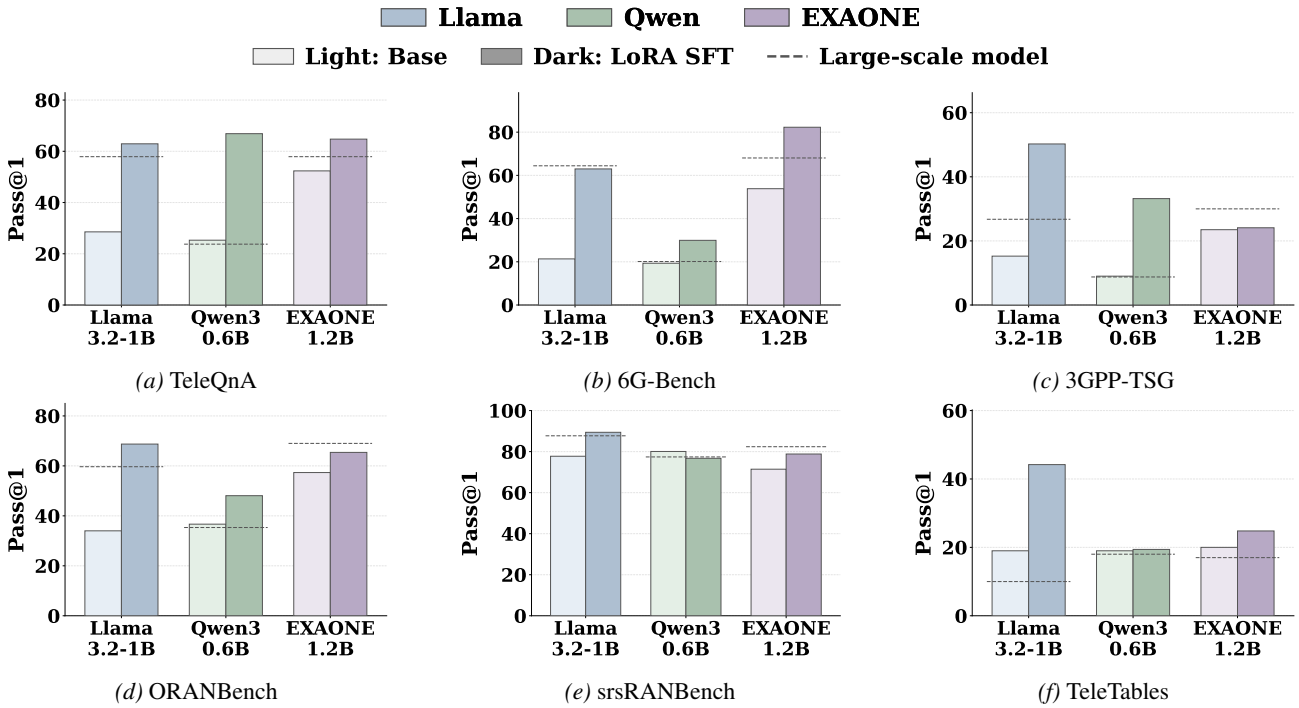
We evaluate three sub-2B SLMs: Llama-3.2-1B (AI@Meta, 2024b), Qwen3-0.6B (Yang et al., 2025), and EXAONE-4.0-1.2B (Bae et al., 2025). These models span three model families – Llama, Qwen, and EXAONE – and are selected for their suitability for edge deployment under strict memory and compute constraints. For comparison, we include the zero-shot performance of large-scale reference models from each corresponding family: Llama-3.1-8B-Instruct (AI@Meta, 2024a), Qwen3-8B (Yang et al., 2025), and EXAONE-3.5-7.8B-Instruct (An et al., 2024). These models serve as a reference point to assess how closely SLMs with fine-tuned LoRA adapters can approach the performance of models several times larger.

3.2. Training Configuration

We apply LoRA with rank $r = 16$ and scaling factor $\alpha = 32$, targeting the query, key, value, output, gate, up-projection, and down-projection matrices. All adapters are trained with a cosine learning rate scheduler with linear warmup and an initial learning rate of 5×10^{-4}, with early stopping based on validation loss. All experiments are conducted on a NVIDIA RTX 3090 with random seed 42.

3.3. Evaluation Protocol

We evaluate each adapter using two metrics. Pass@1 measures whether the model produces the correct answer in a single generation under greedy decoding, and serves as the primary metric throughout our experiments. Maj@k extends this by generating k candidate outputs through temperature sampling and selecting the most frequent answer as the final prediction, offering a more robust performance



(a) TeleQnA *(b)* 6G-Bench *(c)* 3GPP-TSG

(d) ORANBench *(e)* srsRANBench *(f)* TeleTables

Figure 1. Pass@1 comparison across six telecom benchmark categories

estimate when model outputs exhibit variability. We set k as 4 in all sampling-based evaluations. Additionally, we conduct a rank ablation study by varying $r \in \{4, 8, 16, 32\}$ to analyze the trade-off between adapter capacity, memory overhead, and task performance.

4. Results

4.1. Overall Effectiveness across Telecom Tasks

Pass@1 Accuracy Figure 1 presents the Pass@1 of each backbone model before and after fine-tuning, alongside zero-shot performance of large-scale reference models. Across all six task categories and three backbone models, our approach consistently improves accuracy over the zero-shot baseline, demonstrating that LoRA adaptation effectively compensates for limited telecommunications-domain knowledge. The improvement is particularly pronounced in knowledge-intensive categories such as TeleQnA, ORANBench, and srsRANBench. Most notably, in 6G-Bench, EXAONE-4.0-1.2B achieves 79.6%, surpassing its large-scale reference model at 68.1%. In 3GPP-TSG, absolute performance remains lower due to the difficulty of 16-class document classification, but fine-tuning consistently yields improvements. Overall, the performance gap between LoRA-adapted SLMs and large-scale reference models is substantially reduced, supporting the feasibility of compact telecom-specialized models in edge environments.

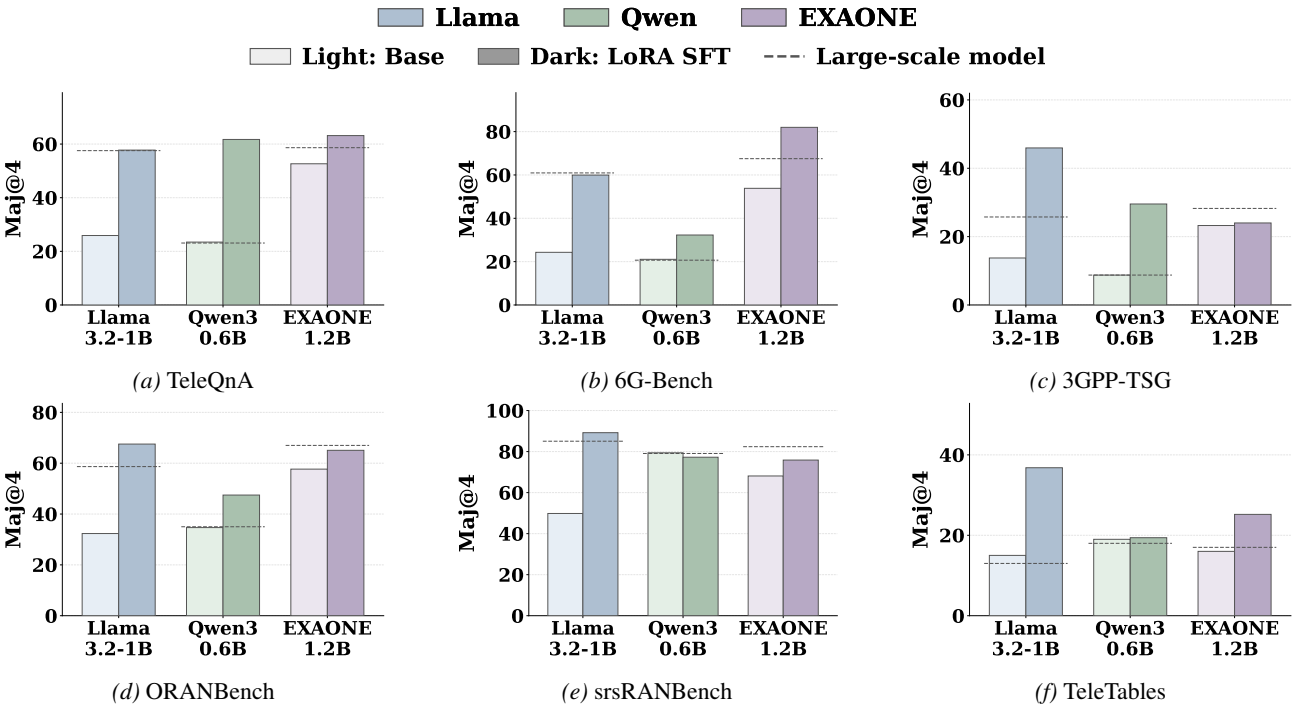
Majority Voting (Maj@4) Figure 2 reports Maj@4 under majority voting over four sampled outputs. Compared with Pass@1 under greedy decoding, Maj@4 yields consistent additional gains across most task categories and backbone models, indicating that the fine-tuned models produce stable and consistent predictions under repeated sampling. The benefit of majority voting is most visible in categories where individual predictions exhibit higher variability, such as ORANBench and 3GPP-TSG, where aggregating multiple outputs helps recover the most consistent prediction. For categories with clearer answer boundaries such as srsRANBench, the gap between Pass@1 and Maj@4 is smaller, suggesting that greedy decoding already captures the dominant prediction in most cases.

Table 2. Comparison of Seen/Unseen performance

	Llama-3.2-1B		**Qwen3-0.6B**		**EXAONE-1.2B**	
Category	Seen	Unseen	Seen	Unseen	Seen	Unseen
TeleQnA	66.38	57.15	72.02	58.15	69.40	56.45
6G-Bench	63.28	61.74	30.14	28.05	83.48	79.60
3GPP_TSG	53.50	45.25	36.33	30.00	26.00	19.00
ORANBench	75.78	61.67	50.67	45.00	65.33	64.33
srsRANBench	92.12	86.05	76.47	78.41	79.36	78.74
TeleTables	58.33	27.00	21.00	12.00	25.67	17.00

4.2. Generalization and Resource Efficiency

Generalization Table 2 compares adapter performance on the training data (seen) and held-out test split (unseen). The performance gap between the two settings remains consistently small across task categories and backbone models,



(a) TeleQnA *(b)* 6G-Bench *(c)* 3GPP-TSG

(d) ORANBench *(e)* srsRANBench *(f)* TeleTables

Figure 2. Maj@4 comparison across six telecom benchmark categories

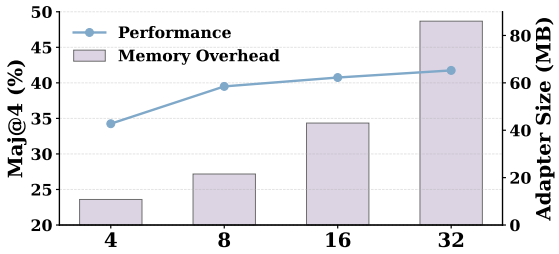


Figure 3. Effect of LoRA rank on Maj@4 evaluated on 3GPP-TSG.

indicating that the fine-tuned adapters generalize to held-out samples within the same benchmark. A particularly interesting case is srsRANBench with Qwen3-0.6B, where unseen performance slightly exceeds seen performance, indicating strong generalization in this category. The limited degradation from seen to unseen evaluation suggests that the adapters learn task-relevant representations rather than simply memorizing training examples, supporting their practical applicability in deployment scenarios where models must handle previously unseen telecommunications queries.

Effect of LoRA Rank Figure 3 analyzes the effect of LoRA rank on task performance and adapter storage size across $r \in \{4, 8, 16, 32\}$, evaluated on 3GPP-TSG using Maj@4 and adapter file size in MB. Performance improves steadily as rank increases from 4 to 8 (34.25% to 41.75%), reflecting the benefit of higher adapter capacity for capturing task-specific patterns. In contrast, increasing the rank from 16 to 32 yields only a marginal gain (40.75% to 41.75%), while doubling the adapter file size from 43.03 MB to 86.03 MB, revealing a clear diminishing return at higher ranks. On 3GPP-TSG, rank 16 provides a favorable balance between performance and storage cost, suggesting a practical default for similar classification-style telecom tasks.

5. Conclusion

This work demonstrates that SLMs equipped with fine-tuned LoRA adapters, despite their compact size, can match or even surpass the performance of models several times larger on telecommunications tasks. Rather than relying on cloud-scale LLMs, compact models can be effectively adapted by updating only a small set of LoRA parameters while keeping the backbone frozen. Experiments on six GSMA ot-full task categories across three sub-2B backbones confirm that LoRA-adapted SLMs consistently improve over their zero-shot baselines and substantially close the gap to larger reference models, highlighting their potential as on-device inference engines for O-RAN and 6G environments where latency, hardware, and privacy constraints preclude cloud-hosted models. Several limitations and future directions remain. A direct comparison with a monolithic adapter trained jointly across all categories would provide a clearer empirical basis for the per-category adaptation strategy. Future work may also explore answer-only loss masking to improve fine-tuning efficiency, adapter merging (Yadav et al., 2023; Yu et al., 2024) to consolidate specialized knowledge into a single generalist adapter, and extension to federated learning settings where adapters are trained locally across distributed edge nodes without sharing raw network data.

References

AI@Meta. Llama 3.1 model card, 2024a. URL https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md.

AI@Meta. Llama 3.2 model card, 2024b. URL https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md.

An, S., Bae, K., Choi, E., Choi, K., Choi, S. J., Hong, S., Hwang, J., Jeon, H., Jo, G. J., Jo, H., et al. Exaone 3.5: Series of large language models for real-world use cases. *arXiv preprint arXiv:2412.04862*, 2024.

Bae, K., Choi, E., Choi, K., Choi, S. J., Choi, Y., Han, K., Hong, S., Hwang, J., Hwang, T., Jang, J., et al. Exaone 4.0: Unified large language models integrating non-reasoning and reasoning modes. *arXiv preprint arXiv:2507.11407*, 2025.

Chen, Z., Zhu, B., Wang, J., Shin, H., Nallanathan, A., and Niyato, D. Network edge inference for large language models: Principles, techniques, and opportunities. *ACM Computing Surveys*, 2026.

Colle, V., Sana, M., Piovesan, N., Domenico, A. D., Ayed, F., and Debbah, M. Telemath: A benchmark for large language models in telecom mathematical problem solving. *IEEE Network*, pp. 1–7, 2026.

Ezzakri, A., Piovesan, N., Sana, M., Domenico, A. D., Ayed, F., and Zhang, H. Teletables: A benchmark for large language models in telecom table interpretation, 2025.

Ferrag, M. A., Lakas, A., and Debbah, M. 6g-bench: An open benchmark for semantic communication and network-level reasoning with foundation models in ai-native 6g networks, 2026.

Gajjar, P. and Shah, V. K. Oran-bench-13k: An open source benchmark for assessing llms in open radio access networks. In *Consumer Communications Networking Conference (CCNC)*, pp. 1–4, 2025a.

Gajjar, P. and Shah, V. K. Oransight-2.0: Foundational llms for o-ran. *IEEE Transactions on Machine Learning in Communications and Networking*, 3:903–920, 2025b.

GSMA. Open telco full benchmarks. https://huggingface.co/datasets/GSMA/ot-full, 2025. Hugging Face dataset, accessed 2026-05-04.

Lu, Z., Li, X., Cai, D., Yi, R., Liu, F., Liu, W., Luan, J., Zhang, X., Lane, N. D., and Xu, M. Demystifying small language models for edge deployment. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 14747–14764, 2025.

Maatouk, A., Ayed, F., Piovesan, N., Domenico, A. D., Debbah, M., and Luo, Z.-Q. Teleqna: A benchmark dataset to assess large language models telecommunications knowledge. *IEEE Network*, 40(2):253–260, 2026.

Mueller, D., Dredze, M., and Andrews, N. Multi-task transfer matters during instruction-tuning. In *Findings of the Association for Computational Linguistics (ACL)*, 2024.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. volume 35, pp. 27730–27744, 2022.

Qu, G., Chen, Q., Wei, W., Lin, Z., Chen, X., and Huang, K. Mobile edge intelligence for large language models: A contemporary survey. *IEEE Communications Surveys Tutorials*, 27(6):3820–3860, 2025.

Sana, M., Piovesan, N., Domenico, A. D., Kang, Y., Zhang, H., Debbah, M., and Ayed, F. Reasoning language models for root cause analysis in 5g wireless networks, 2025.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners, 2021.

Yadav, P., Tam, D., Choshen, L., Raffel, C. A., and Bansal, M. Ties-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 7093–7115, 2023.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yin, X., You, F., Du, H., and Huang, K. Ubiquitous intelligence via wireless network-driven llms evolution. *npj Wireless Technology*, 2(1):4, 2026.

Yu, L., Yu, B., Yu, H., Huang, F., and Li, Y. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *International Conference on Machine Learning (ICML)*, 2024.

Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., and Finn, C. Gradient surgery for multi-task learning. volume 33, pp. 5824–5836, 2020.

Zhou, H., Hu, C., Yuan, Y., Cui, Y., Jin, Y., Chen, C., Wu, H., Yuan, D., Jiang, L., Wu, D., et al. Large language model (llm) for telecommunications: A comprehensive survey on principles, key techniques, and opportunities. *IEEE Communications Surveys & Tutorials*, 27(3):1955–2005, 2024.

Zhou, Y., Wang, T., Wu, Y., Cai, P., Zhou, F., and Shi, Y. Generative and large ai models for 6g wireless networks: The optimization perspective. *Engineering*, 2026.

Zou, H., Zhao, Q., Tian, Y., Bariah, L., Bader, F., Lestable, T., and Debbah, M. Telecomgpt: A framework to build telecom-specific large language models. *IEEE Transactions on Machine Learning in Communications and Networking*, 3:948–975, 2025.

Zou, H., Zhao, Q., Lasaulce, S., Zhang, C., Tian, Y., Bariah, L., Bader, F., and Debbah, M. Large language models in 6g from standard to on-device networks. *Nature Reviews Electrical Engineering*, pp. 1–12, 2026.