Evaluating Open-Source Large Language Models for Technical Telecom Question Answering

Arina Caraus*†, Alessio Buscemi*, Sumit Kumar*, and Ion Turcanu*
*Luxembourg Institute of Science and Technology (LIST), Luxembourg
†RMT Labs, Luxembourg
acaraus@rmt-labs.com { alessio.buscemi, sumit.kumar, ion.turcanu }@list.lu

***Abstract*—Large Language Models (LLMs) have shown remarkable capabilities across various fields. However, their performance in technical domains such as telecommunications remains underexplored. This paper evaluates two open-source LLMs, Gemma 3 27B and DeepSeek R1 32B, on factual and reasoning-based questions derived from advanced wireless communications material. We construct a benchmark of 105 question–answer pairs and assess performance using lexical metrics, semantic similarity, and LLM-as-a-judge scoring. We also analyze consistency, judgment reliability, and hallucination through source attribution and score variance. Results show that Gemma excels in semantic fidelity and LLM-rated correctness, while DeepSeek demonstrates slightly higher lexical consistency. Additional findings highlight current limitations in telecom applications and the need for domain-adapted models to support trustworthy Artificial Intelligence (AI) assistants in engineering.**

***Index Terms*—Large Language Models, Telecommunications, Artificial Intelligence**

I. Introduction

Large Language Models (LLMs) are sophisticated Artificial Intelligence (AI) models whose popularity is rapidly increasing, with successful applications already demonstrated in fields such as psychology [1], medicine [2], law [3], and finance [4]. Despite this widespread adoption, their use in the telecommunications sector remains relatively limited. Nonetheless, recent efforts have emerged to develop domain-specific LLMs aimed at enhancing tasks such as customer service automation, dynamic network configuration, and traffic classification. The limited integration of LLMs in telecommunications can be attributed to the domain's inherent complexity. Unlike applications that rely primarily on linguistic and contextual reasoning, telecom tasks demand a solid understanding of physical and mathematical principles, such as electromagnetic wave propagation and modulation schemes. Effective deployment in this space requires models to interpret formal systems, follow protocol specifications, and reason over signal processing operations. These demands stretch the capabilities of current general-purpose models and underscore the need for specialized solutions tailored to the domain's unique requirements.

Although recent LLM families – such as GPT, Llama, DeepSeek, Gemma, and Mistral – have shown rapid progress, their ability to process highly technical, domain-specific content in telecommunications remains largely untested. In particular, it is unclear whether such models can produce accurate and meaningful responses to questions requiring deterministic reasoning, structured derivations, and textbook-based knowledge.

In this paper, we aim to fill this gap by constructing and using a rigorous evaluation benchmark derived from the textbook *Wireless Communications* by Molisch [5], a foundational reference in both academic and industrial settings. Verified answers are sourced from the textbook's official solution manual to ensure high fidelity and precision. The resulting dataset provides a robust testbed for measuring LLM performance on factual Question-Answer (QA) in the telecommunications domain. The goal of this work is to assess whether LLMs can act as reliable assistants for telecom engineers and operators.

In addition, trust is a fundamental prerequisite. Knowing that a model can produce correct answers in principle is not sufficient: if its outputs are inconsistent across repetitions of the same question, the user may unknowingly rely on flawed information, potentially leading to incorrect technical decisions. For this reason, we place a strong emphasis not only on the *correctness* of the answers, but also on the *consistency* with which models provide them. This consistency is measured both in answer generation and in the evaluation process when models serve as judges of quality.

Our study is guided by the following research questions:

- **RQ1:** Can LLMs provide technically accurate and complete answers to telecommunications questions?
- **RQ2:** What metrics best capture the quality and utility of the answers generated by LLMs?
- **RQ3:** How consistent are open-source LLMs in both generating answers and evaluating answer quality when exposed to the same technical questions in the telecommunications domain?

To address these questions, we develop a structured evaluation framework. For RQ1, we assess whether LLMs can produce technically accurate and complete responses to questions covering diverse areas of telecommunications. For RQ2, we evaluate answer quality using a combination of metrics: lexical similarity, semantic similarity, and assessments provided by a separate LLM acting as a judge. RQ3 is addressed by analyzing how consistent the models are in their generated responses and in the evaluations they provide when presented with identical technical prompts. This study provides one of the first comprehensive assessments of state-of-the-art LLMs on technical telecommunications content across multiple areas.

IEEE Global Communications Conference (GLOBECOM 2025)

II. Related Work

Recent research has explored the capabilities and limitations of LLMs in both general-purpose and domain-specific contexts. Zhou et al. [6] and Maatouk et al. [7] point out that prompting techniques such as In-Context Learning (ICL) and Chain-of-Thought (CoT) often fall short in evaluate complex tasks such as multi-hop reasoning and factual consistency. General benchmarks like HellaSwag, SuperGLUE, and MMLU assess overall NLP performance but lack the granularity needed for expert-level knowledge in specialized fields like telecommunications.

In the telecom domain, Maatouk et al. [8] introduced the TeleQnA dataset, evaluating models such as GPT-3.5 and GPT-4 with multiple-choice and open-ended questions from 3GPP standards and telecom-specific lexicons. Soman and Ranjani [9] focus on classifying technical documents from 3GPP's TSGs (RAN, SA, and CT), underscoring the deep technical expertise required for accurate categorization. However, most of these efforts concentrate heavily on 3GPP materials and use primarily multiple-choice formats, limiting their breadth [10].

Bariah et al. [11] emphasize the growing importance of explainability in telecom-specific LLMs, especially as these models are integrated into network operations and decision-making systems. Despite the utility of datasets like TeleQnA [8], recent literature suggests that existing benchmarks do not fully capture the diversity of telecom knowledge [6]. Ahmed et al. [12], for example, employed this benchmark to evaluate the performance of GPT-3.5 models on various telecom tasks.

Another emerging area of research is the development of conversational assistants tailored to telecom applications [13]. These systems demand not only accuracy but also contextual understanding and dialog-level reasoning.

Our work expands upon prior efforts in three key ways: (1) we evaluate the lesser-studied DeepSeek model and compare its performance with Gemma; (2) we shift the focus from multiple-choice to free-form question answering; and (3) our dataset encompasses a broader spectrum of telecom topics beyond 3GPP, offering a more diverse and realistic benchmark for assessing LLM performance in this domain.

III. Methodology

A. *Dataset Construction*

We built a structured QA dataset based on the textbook "Wireless Communications" (2nd ed.) by Molisch [5], chosen for its technical depth and domain relevance. The dataset includes 105 questions, with $\approx 60\,\%$ being conceptual and $\approx 40\,\%$ requiring mathematical derivations or calculations. Reference answers were extracted from the official Solution Manual for "Exercises in the Textbook Wireless Communication by A.F. Molisch" [14], providing high-quality ground truth annotations.

Each data instance is formatted as a JSON object organized by chapter, and includes three fields: `question` (the original exercise from the textbook), `answer` (the corresponding detailed solution from the manual), and `final_answer` (a concise, manually curated summary that highlights key facts or numerical values). The `final_answer` field was specifically introduced to support evaluation tasks and to better align with the output structure of LLMs.

To ensure consistency, we retained only QA pairs with both the `answer` and `final_answer` in the solutions manual. This filtering yielded a refined dataset of 105 high-quality QA pairs spanning multiple chapters. The resulting dataset provides a domain-specific benchmark for assessing the factual performance of LLMs in the context of wireless communications. An example schema is shown in Figure 1.

B. *Evaluation Framework*

To systematically assess the performance of LLMs on domain-specific technical questions, we implemented a multi-tiered evaluation framework that integrates both automatic metrics and model-based judgments. As illustrated in Figure 2, the framework operates at the level of individual QA pairs

JSON Template

```
{"chapter": "<Chapter number and title>",
"questions":
[
{"question": "<Original book question>",
 "answer": "<Full solution text>",
 "final_answer": "<Short target answer>"},
...
]}
```

Figure 1. Test Dataset JSON Template. This format defines the structure used to evaluate an LLM's performance. It includes a set of chapter-based questions, full solution texts, and expected short-form answers (used as ground truth) to compare against generated outputs.

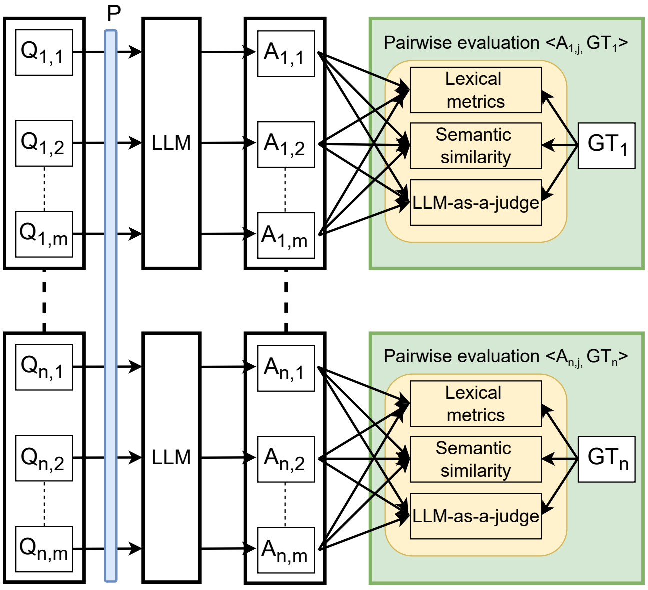


Figure 2. Evaluation pipeline: each question Q_i is submitted m times to the LLM, yielding answers $A_{i,1}, \dots, A_{i,m}$, each evaluated against the ground truth GT_i using lexical, semantic, and judgment-based metrics.

and applies three complementary evaluation strategies: lexical similarity, semantic similarity, and LLM-as-a-Judge scoring.

Each question Q_i in the dataset is submitted m times to an LLM L using a structured prompt P, described in Section III-C. In this work, we selected $n = 105$ questions, as described in Section III-A, and set $m = 20$ repetitions per question. This configuration was chosen to balance computational cost and statistical robustness, and to enable a consistency analysis of the model's behavior across repeated queries. The resulting answers $A_{i,1}, \dots, A_{i,m}$ are compared with the reference answer – i.e., Ground Truth (GT) – through pairwise evaluation.

Lexical similarity is assessed using BLEU, ROUGE-L, and METEOR metrics, which capture surface-level token overlap but may penalize legitimate variation in expression. Semantic similarity is computed via cosine distance between sentence embeddings of $A_{i,j}$ and GT_i, offering a meaning-oriented perspective on answer quality. Finally, in the *LLM-as-a-Judge* setup, a separate language model is prompted to evaluate each answer by directly comparing it with the ground truth GT_i. It assigns a score from 0 to 10, where 0 indicates a response that is entirely incorrect or irrelevant, and 10 corresponds to a perfectly aligned and factually accurate answer.

In addition to answer quality, we examine the reliability of LLMs through two forms of consistency. Answer consistency captures the variability in responses across the m independent generations of the same question, offering insight into the model's determinism. Scoring consistency reflects the variation in repeated judgment-based evaluations of identical answers. Together, these metrics provide a robust and interpretable assessment of both performance and behavioral stability – key requirements for telecom-grade applications.

C. Prompting

In this work, we explore two complementary prompting setups, *zero-shot* and *few-shot*, to examine the responses generated by the LLM. In the zero-shot configuration, the model is presented solely with the raw textbook question Q_i and is expected to produce an answer without any additional guidance or contextual examples. An example of such an input is shown in Figure 3.

The few-shot setup follows a different approach. Here, the same question Q_i is embedded in a structured instruction template, shown in Figure 4, which includes two worked examples from the telecommunications domain to provide context and guidance. The prompt also explicitly requires the model to generate only a concise `final_answer`, and mandates citation of the source material. To enhance response diversity and enable a more robust evaluation, we query the model $m = 20$ times using this few-shot prompt, resulting in a response set $\{R^{\text{LLM}}_{i,1}, \dots, R^{\text{LLM}}_{i,m}\}$. This ensemble of answers serves as the foundation for the pairwise and ground truth similarity analysis described in Section III.

Zero-shot Input

```
Assuming that directions of arrival are uniformly
distributed at the MS, how large is the
correlation coefficient (for a GSM-1800 system)
between the channel in the middle and the end of
the burst when the MS moves at 250~km/h? How
large is the correlation coefficient between the
channels at the beginning and end of the burst?
```

Figure 3. An example of zero-shot input

D. Additional details

To assess the transparency and potential hallucination behavior of the LLM when answering technical questions,

Few-shot Prompt Template

```
You are given a question related to
telecommunications, networking, or signal
transmission.
Your task is:
1. Analyse the question carefully.
2. Provide only the final_answer with no
additional explanation or reasoning.
3. The final_answer must be concise, accurate,
and directly respond to the question.
4. At the end give the source from where you
obtained the information. Use the format-
Source:

Examples
--------
Input Question: Which of the following systems
cannot transmit in both directions (duplex or
semiduplex): (i) cellphone, (ii) cordless
phone, (iii) pager, (iv) trunking radio, (v)
TV broadcast system?
Output: pager, TV broadcast

Input Question: Communication is to take place
from one side of a building to the other as
depicted in Figure 30.1, using 2-m-tall
antennas.
Convert the building into a series of
semi-infinite screens and determine the field
strength at the receive antenna caused by
diffraction using Bullington's method for (a)
f = 900 MHz, (b) f = 1 800 MHz, and (c) f =
2.4 GHz.
Output: (a) 0.0155, (b) 0.0109, (c) 0.0095

Input Question: {question_text}
```

Figure 4. An example of a few-shot prompt template

Evaluation JSON Template

```
{
"question": "<QA input>",
"generated_answer": "<LLM reply>",
"rating": "<Judge LLM comment>",
"rating_value": <Judge score>,
"source": "<LLM source or 'NaN'>",
}
```

Figure 5. An example of an evaluation JSON template

we incorporate a mandatory source attribution component into the prompt. Each prompt instructs the LLM to append a source for the answer it provides.

Regarding the LLM-as-a-Judge evaluation, we apply a rule-based outlier filtering mechanism focused solely on enforcing rating boundaries to ensure the validity of the evaluation scores. Let $\mathcal{S} = \{s_{i,j}\}$ ($i \in [1, n]$, $j \in [1, m]$) denote the set of scores assigned by the LLM-as-Judge to the generated answers. Each score s_i must fall within the interval $[0, 10]$, as specified in the scoring instructions. All scores s_i that violate this constraint are discarded from further analysis. This step is essential to ensure that metrics and aggregations reflect only interpretable and rule-compliant scores. For the evaluation with LLM-as-a-Judge, the data must be in the template shown in Figure 5.

IV. Performance Evaluation

A. Experimental setup

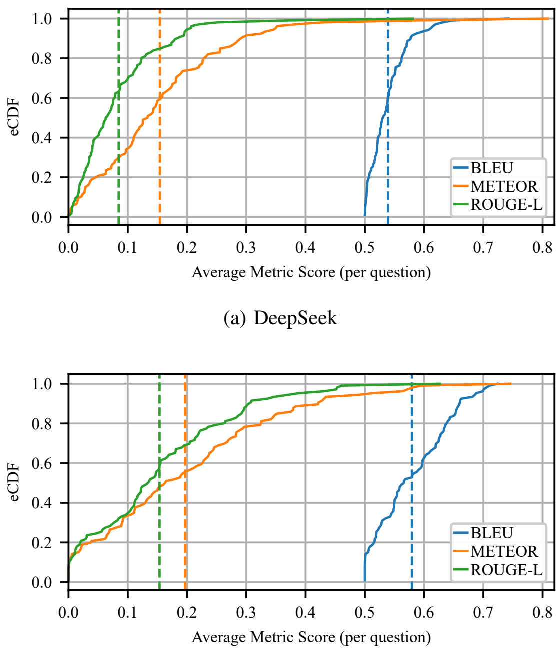
In this work, we use two advanced language models with complementary reasoning capabilities. DeepSeek-R1 32B is a reasoning-focused model trained to generate explicit, multi-step rationales, making it suitable for tasks requiring logical transparency. Gemma 3 27B, by contrast, is a lightweight instruction-tuned model optimized for efficient, coherent responses in general-purpose applications. Both models were deployed locally on a single Nvidia Tesla V100 (32 GByte) GPU using the Ollama library in Python, and were pulled from Hugging Face. All processing and evaluation was conducted using Python 3.11.

Before conducting the performance evaluation, we first verified whether the models had prior exposure to the textbook solutions used in our test set. This step is essential to determine whether the models generate answers by reasoning through the problem or simply by recalling content memorized during training. To this end, we selected a subset of questions and prompted each model multiple times, explicitly asking for the sources of their answers, as described in Section III-C.

For DeepSeek-R1, none of the answers cited the textbook or its solutions manual. Instead, references were consistently made to other general telecommunications literature [15, 16, 17]. This indicates that DeepSeek-R1 was likely not exposed to the ground-truth answers during training. By contrast, Gemma 3 27B occasionally cited the textbook from which the questions were drawn, but never mentioned the corresponding solutions manual. However, it frequently referenced other works [18, 19]. From this, we conclude that while Gemma may have encountered the source of the questions, neither model had access to the actual answers. This gives us confidence that their performance reflects genuine reasoning ability rather than memorization.

B. Accuracy analysis

The accuracy of answers generated by the LLM is calculated taking into consideration the correctness of reasoning to find the right answer. We analyze whether the reasoning followed by the LLM aligns with the GT. To complement this analysis, we evaluate the similarity between the generated and reference answers at both lexical and semantic levels.



(a) DeepSeek

(b) Gemma

Figure 6. Accuracy analysis of answers A_i generated by LLM with Gemma and DeepSeek with respect to GT for questions Q_i.

1) Lexical: We used normalized BLEU-4, ROUGE-L, and METEOR metrics for lexical similarity. These metrics measure word and phrase overlap between candidate and GT texts. They focus on precision, recall, and matching n-grams, rewarding outputs that closely follow the reference structure.

Figure 6 shows that Gemma consistently achieves higher scores than DeepSeek across all three metrics. Its score distributions are shifted toward the right, indicating a greater proportion of high-quality outputs. For example, around 40 % of Gemma's BLEU scores exceed 0.6, compared to only 5 % for DeepSeek. Similarly, more than 20 % of Gemma's METEOR scores surpass 0.3, while only 10 % of DeepSeek's scores surpass this threshold. These results indicate that Gemma produces more precise, fluent, and structurally aligned responses, outperforming DeepSeek in both lexical and semantic terms.

2) Semantic: For semantic similarity, we use cosine similarity computed over sentence embeddings to assess whether the meaning conveyed by the generated answer is close to that of the GT, even if the wording differs. Specifically, we use the all-MiniLM-L6-v2 model to produce the embeddings. This allows us to evaluate the deeper alignment in content beyond surface-level token matching. The results illustrated in Figure 7 show that Gemma provides better performance in terms of semantic similarity, achieving an average cosine similarity of 0.4, compared to 0.3 achieved by DeepSeek.

3) LLM-as-a-Judge: To complement the automatic evaluation, we also considered LLM-as-a-Judge scores, where the same LLM is used to both answer the question and evaluate the accuracy of the response. The results, shown in Figure 8, indicate that Gemma again outperforms DeepSeek, with a greater proportion of its responses receiving higher ratings. This subjective evaluation aligns well with the automatic metrics, confirming that Gemma's outputs are not only more lexically and structurally aligned with references, but also judged as more accurate and relevant by a powerful language model.

C. Consistency analysis

To assess the consistency of the LLM when answering the same question multiple times, we repeated each question Q_i exactly $m = 20$ times. We then computed the semantic and lexical similarity between each pair of consecutive responses, $\langle A_{i,j}, A_{i,j+1}\rangle$ for $j \in [1, m-1]$. In addition, we evaluated variability by calculating the standard deviation of all 20 responses $A_{i,j}$ for each Q_i. For the LLM-as-a-Judge, we similarly measured consistency by computing the standard deviation of the scores it assigned across multiple judgments of the same answer.

1) Lexical: To assess lexical and structural consistency, we measured the standard deviation of BLEU, METEOR, and ROUGE-L scores between consecutive outputs generated in response to the same prompt. Figure 9 reveals that DeepSeek is consistently more stable than Gemma across all three metrics. For example, over 70 % of DeepSeek's BLEU pairwise deviations fall below 0.1, whereas Gemma's BLEU standard deviations are more broadly distributed, with many values exceeding 0.15. The same pattern holds for ROUGE-L and METEOR, where DeepSeek's curve rises steeply, indicating that a greater proportion of its outputs vary less between repeated runs. These findings suggest that while Gemma may produce higher average-quality responses, DeepSeek delivers more reproducible outputs – a valuable trait in applications where deterministic behavior and response stability are critical.

2) Semantic: Figure 10 compares the standard deviation of cosine similarity scores between consecutive outputs for the same question. As it can be noted, Gemma shows higher semantic consistency than DeepSeek, indicating more stable and deterministic outputs with less variability in meaning when answering the same prompt multiple times.

3) LLM-as-a-Judge: The consistency of the LLM-as-a-Judge was evaluated by measuring the standard deviation of the scores it assigned across repeated evaluations of the same question. Figure 11 shows that Gemma receives more consistent judgments than DeepSeek. Specifically, over 80 %

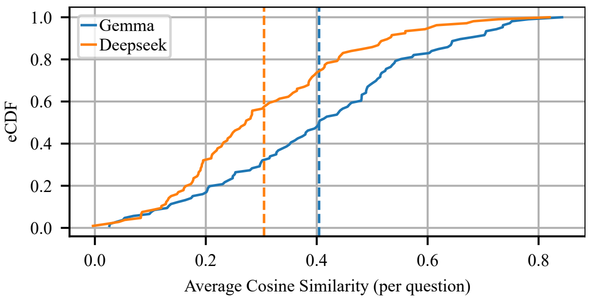


Figure 7. Accuracy analysis of answers A_i generated by LLM with Gemma and DeepSeek using cosine similarity.

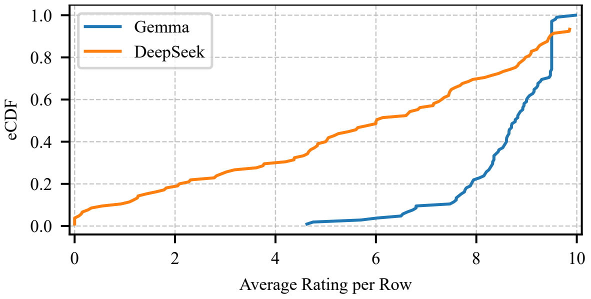
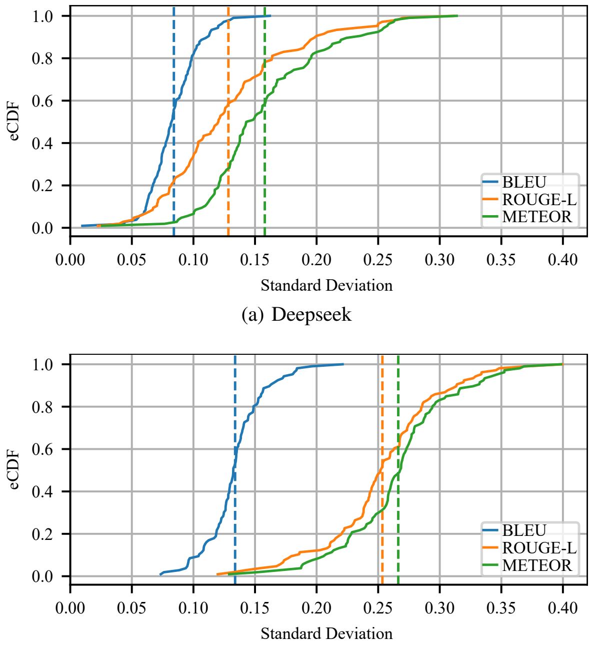


Figure 8. Accuracy analysis of answers A_i generated by LLM with Gemma and DeepSeek using LLM-as-a-Judge for scoring.



(a) Deepseek

(b) Gemma

Figure 9. Lexical consistency analysis of answers generated by LLM with Gemma and DeepSeek calculated on pairs of answers $\langle A_i^j, A_i^{j+1}\rangle \forall Q_i$.

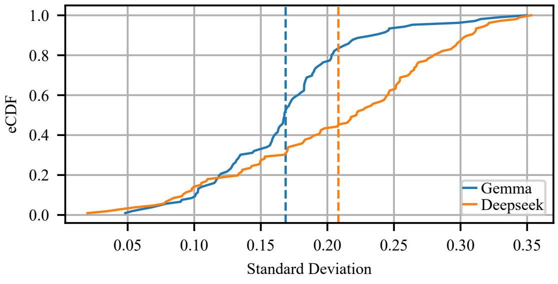


Figure 10. Semantic consistency analysis of answers A_i generated by LLM using Gemma and DeepSeek.

of Gemma's judgment standard deviations fall below 1.0, whereas DeepSeek's curve rises more slowly, indicating greater variability in its scores. This suggests that the LLM-as-a-Judge is more stable and confident when evaluating Gemma's outputs. A likely explanation is that Gemma's responses maintain a more consistent level of quality and clarity, leading to fewer ambiguities in assessment, while DeepSeek's more variable output quality results in fluctuating judgments.

V. Conclusion

This paper presented a comprehensive evaluation of two state-of-the-art open-source Large Language Models (LLMs) – Gemma 3 27B and DeepSeek R1 32B – on a set of challenging, factual and reasoning-based questions derived from advanced wireless communications. Our goal was to assess their ability to serve as reliable AI assistants in the telecommunications domain, where correctness and consistency are essential. By constructing a rigorous benchmark from authoritative textbook material, we were able to assess the quality and the stability of the models' outputs using a mix of lexical metrics, semantic similarity, and subjective scoring through LLM-as-a-Judge.

Our results reveal a nuanced performance landscape. Gemma consistently outperforms DeepSeek in terms of semantic fidelity, lexical precision, and LLM-rated accuracy, producing more fluent and technically relevant responses. However, when it comes to lexical consistency, i.e, the ability to reproduce similar answers across multiple generations, DeepSeek proves to be more stable, with lower variation in BLEU, METEOR, and ROUGE-L scores. Interestingly, Gemma's outputs receive more consistent evaluations from the LLM-as-a-Judge, indicating higher clarity and less ambiguity in how its answers are perceived. Together, these findings suggest a trade-off between raw answer quality and output determinism, highlighting the importance of evaluating both correctness and consistency for technical deployments.

Future work will focus on expanding the benchmark to include questions related to 5G and 6G standards, thereby broadening the coverage of real-world telecommunications scenarios. In addition, we plan to evaluate closed-source models, such as GPT-4 and Claude, within the same framework to provide a more complete view of current model capabilities. Finally, we aim to incorporate human-in-the-loop feedback mechanisms to support iterative refinement and trust calibration, enabling more reliable deployment of LLMs in high-stakes engineering tasks. Looking further ahead, we also see potential in evaluating LLMs within agent-based paradigms, where techniques such as Chain-of-Thought (CoT) and Retrieval-Augmented Generation (RAG) could play a central role.

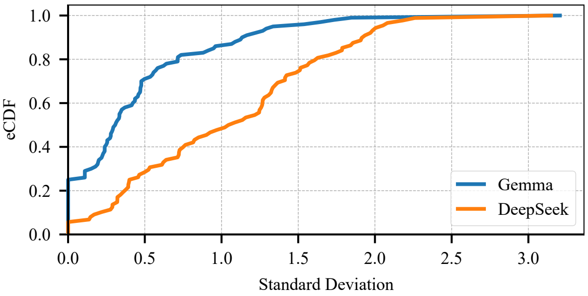


Figure 11. Comparison of Gemma and DeepSeek model performance, showing the Empirical Cumulative Distribution Function (eCDF) of Standard Deviation.

References

[1] J. Hu, T. Dong, L. Gang, H. Ma, P. Zou, X. Sun, D. Guo, X. Yang, and M. Wang, "PsycoLLM: Enhancing LLM for Psychological Understanding and Evaluation," *IEEE Transactions on Computational Social Systems*, vol. 12, no. 2, pp. 539–551, 2024.

[2] Q. Xie, Q. Chen, A. Chen, C. Peng, Y. Hu, F. Lin, X. Peng, J. Huang, J. Zhang, V. Keloth, et al., "Me-LLaMA: Foundation Large Language Models for Medical Applications," *Research square*, 2024.

[3] M. Siino, M. Falco, D. Croce, and P. Rosso, "Exploring LLMs Applications in Law: A Literature Review on Current Legal NLP Approaches," *IEEE Access*, vol. 13, pp. 18 253–18 276, 2025.

[4] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, "BloombergGPT: A Large Language Model for Finance," *arXiv preprint arXiv:2303.17564*, 2023.

[5] A. F. Molisch, *Wireless Communications*, 2nd ed. Chichester, UK: John Wiley & Sons, 2011.

[6] H. Zhou, C. Hu, Y. Yuan, Y. Cui, Y. Jin, C. Chen, H. Wu, D. Yuan, L. Jiang, D. Wu, et al., "Large Language Model (LLM) for Telecommunications: A Comprehensive Survey on Principles, Key Techniques, and Opportunities," *IEEE Communications Surveys & Tutorials*, vol. 27, no. 3, pp. 1955–2005, 2024.

[7] A. Maatouk, N. Piovesan, F. Ayed, A. De Domenico, and M. Debbah, "Large Language Models for Telecom: Forthcoming Impact on the Industry," *IEEE Communications Magazine*, vol. 63, no. 1, pp. 62–68, 2024.

[8] A. Maatouk, F. Ayed, N. Piovesan, A. De Domenico, M. Debbah, and Z.-Q. Luo, "TeleQnA: A Benchmark Dataset to Assess Large Language Models Telecommunications Knowledge," *IEEE Network*, 2025.

[9] S. Soman and H. Ranjani, "Observations on LLMs for Telecom Domain: Capabilities and Limitations," in *Proceedings of the Third International Conference on AI-ML Systems*, 2023, pp. 1–5.

[10] A.-L. Bornea, F. Ayed, A. De Domenico, N. Piovesan, and A. Maatouk, "Telco-RAG: Navigating the Challenges of Retrieval Augmented Language Models for Telecommunications," in *IEEE Global Communications Conference*, IEEE, 2024, pp. 2359–2364.

[11] L. Bariah, Q. Zhao, H. Zou, Y. Tian, F. Bader, and M. Debbah, "Large Generative AI Models for Telecom: The Next Big Thing?" *IEEE Communications Magazine*, vol. 62, no. 11, pp. 84–90, 2024.

[12] T. Ahmed, N. Piovesan, A. De Domenico, and S. Choudhury, "Linguistic Intelligence in Large Language Models for Telecommunications," in *IEEE International Conference on Communications Workshops (ICC Workshops)*, IEEE, 2024, pp. 1237–1243.

[13] L. Bariah, H. Zou, Q. Zhao, B. Mouhouche, F. Bader, and M. Debbah, "Understanding Telecom Language Through Large Language Models," in *IEEE Global Communications Conference*, IEEE, 2023, pp. 6542–6547.

[14] P. Almers, O. Edfors, F. Florén, A. Johanson, J. Karedal, B. K. Lau, A. F. Molisch, A. Stranne, F. Tufvesson, S. Wyne, and H. Zhang, *Solution Manual for the Exercises in the Textbook Wireless Communications by A. F. Molisch*. John Wiley and Sons, Ltd.

[15] S. Haykin, *Communication systems*. John Wiley & Sons, 2008.

[16] J. G. Proakis, M. Salehi, N. Zhou, and X. Li, *Communication systems engineering*. Prentice Hall New Jersey, 1994, vol. 2.

[17] R. E. Ziemer and W. H. Tranter, *Principles of communications*. John Wiley & Sons, 2014.

[18] J. G. Proakis and M. Salehi, *Digital communications*. McGraw-hill New York, 2001, vol. 4.

[19] T. S. Rappaport, *Wireless communications: Principles and practice, 2/E*. Pearson Education India, 2010.