Advancing Telecoms with Large Language Models: Benchmarks, Datasets, and Innovations

Antonio De Domenico
Huawei Technologies

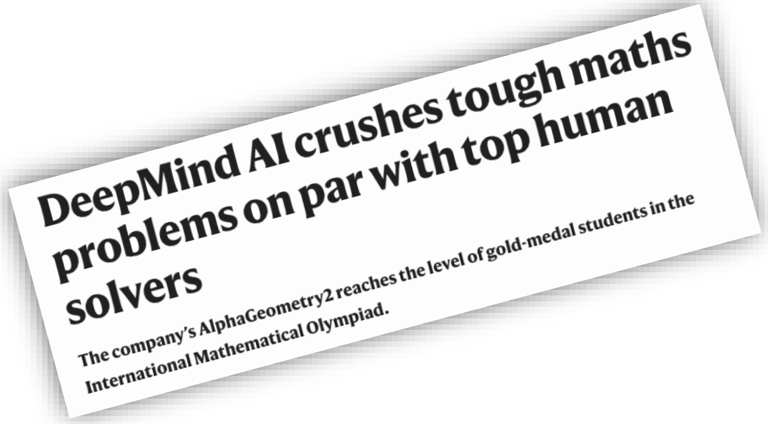
AI for Good
Global Summit

Identifying innovative AI applications to solve global challenges

HUAWEI

The LLM Effect: Threat or Catalyst for Telecom Growth?

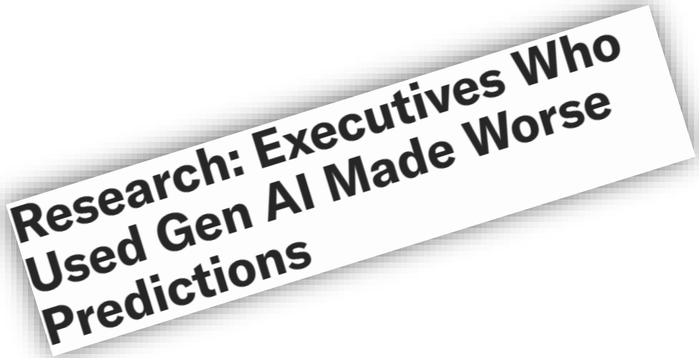
🧠 **Mathematics Olympiad =** ✅



DeepMind AI crushes tough maths problems on par with top human solvers

The company's AlphaGeometry2 reaches the level of gold-medal students in the International Mathematical Olympiad.

📲 **Executive decision co-pilot =** ❌



Research: Executives Who Used Gen AI Made Worse Predictions

📲 **Telecoms =** ❌

🔍 **1. Poor Accuracy on Technical Queries**

Test: GPT-4 & Claude vs. human experts on 'RAN slicing vs. Network slicing'

- ❌ 30-40% incorrect responses (Confused 3GPP standards)

📜 **2. Hallucinations in Regulations & Standards**

Test: Asked AI about 5G spectrum policies

❌ Invented non-existent frequency bands & misquoted ITU rules

⚡ **3. Inadequate for Network Troubleshooting**

Real-world trial: AI suggested fixes that would have worsened packet loss

❌ Failed to interpret vendor-specific network KPIs (SNR, RRC success rate)

💡 **Why it matters: telcos risk wasted investments, regulatory risks, poor user experience, and falling behind AI-native competitors.**

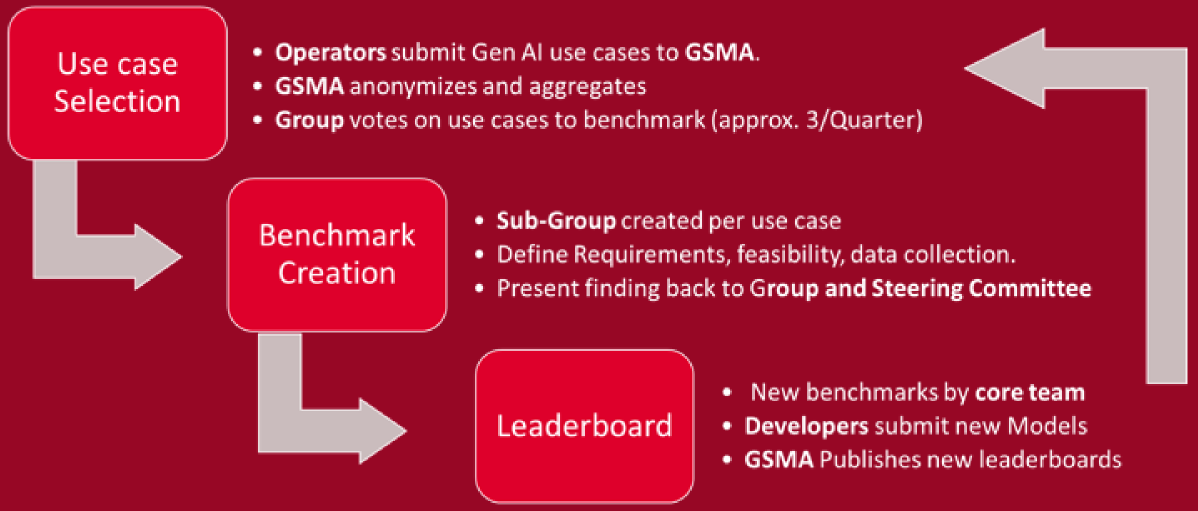
https://arxiv.org/pdf/2407.09424
https://aclanthology.org/2024.emnlp-industry.45.pdf
https://hbr.org/2025/07/research-executives-who-used-gen-ai-made-worse-predictions?ab=HP-hero-latest-1

Open-Telco LLM Benchmarks

The open-source hub of Telco specific models, data and evaluation

GSMA Open-Telco LLM Benchmarks

Open-source community aimed at understanding and improving the performance of large language models (LLMs) for telecom-specific applications.



Use case Selection

- **Operators** submit Gen AI use cases to **GSMA**.
- **GSMA** anonymizes and aggregates
- **Group** votes on use cases to benchmark (approx. 3/Quarter)

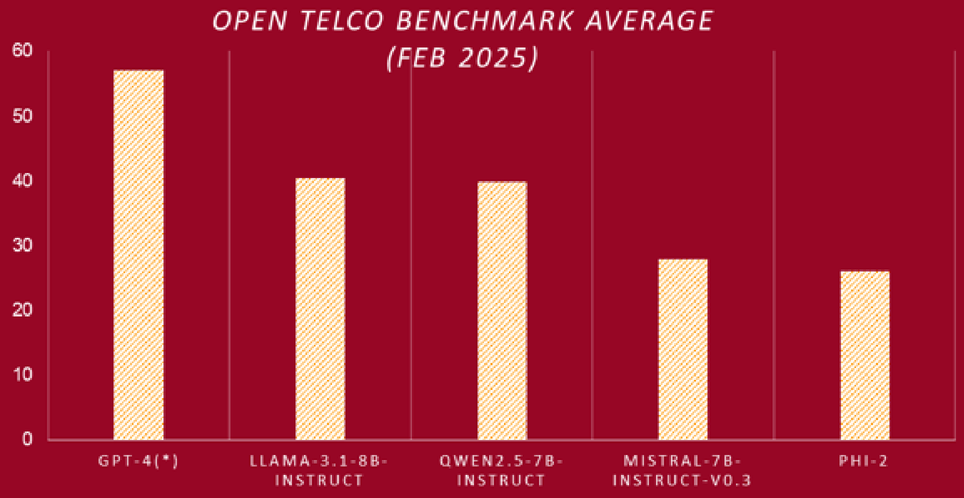
Benchmark Creation

- **Sub-Group** created per use case
- Define Requirements, feasibility, data collection.
- Present finding back to **Group and Steering Committee**

Leaderboard

- New benchmarks by **core team**
- **Developers** submit new Models
- **GSMA** Publishes new leaderboards

Holistic Evaluation on Telecom, Maths, Logic…

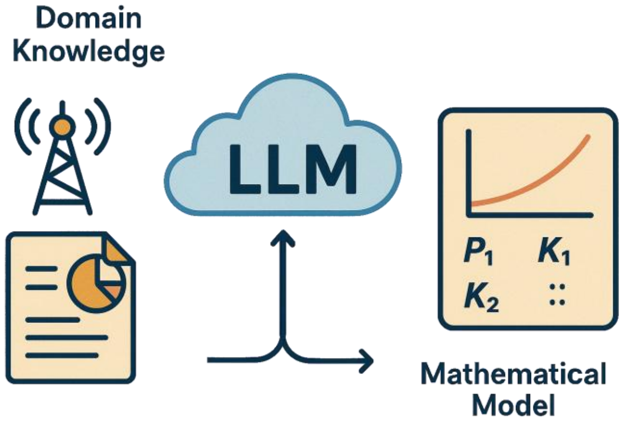


GSMA Open-Telco LLM Benchmarks – Blog post - https://huggingface.co/blog/otellm/gsma-benchmarks
GSMA Open-Telco LLM Benchmarks – Leaderboard - https://huggingface.co/spaces/otellm/open-telecom-llm-leaderboard

TeleFamily Benchmark: from Telecom Knowledge to the Driving Seat of Network M&O

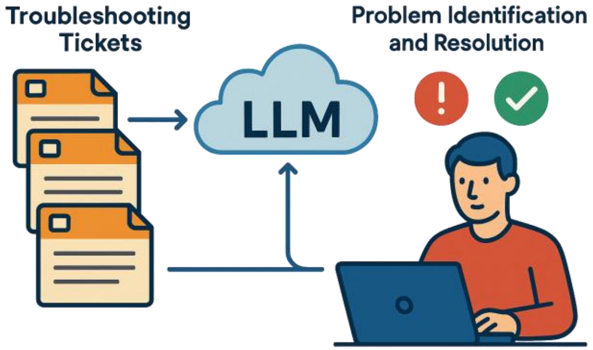
TeleQnA

10000 multiple choice questions aiming to test LLM knowledge on telecom networks, from research to standard



TeleMath

500 curated QnAs to test LLMs on solving problems with numerical solutions within the telecom domain

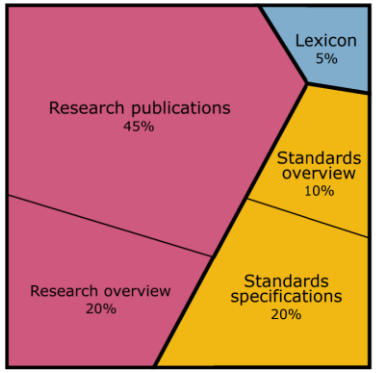


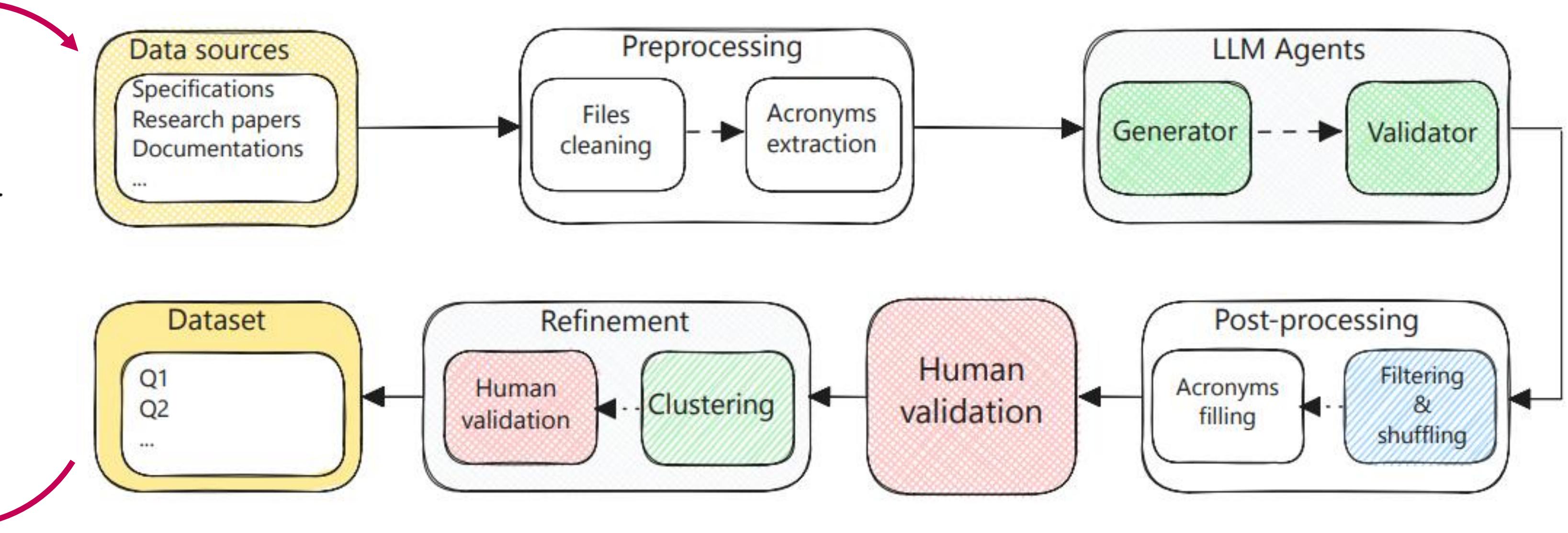
TeleM&O

New datasets to test agentic capabilities, troubleshooting, policy management…

https://huggingface.co/datasets/netop/TeleQnA
https://huggingface.co/datasets/netop/TeleMath

How can we Evaluate the Telecoms Knowledge of a LLM?





Question 5: What is the maximum number of eigenmodes that the MIMO channel can support? (nt is the number of transmit antennas, nr is the number of receive antennas)

- *Option 1:* nt
- *Option 2:* nr
- *Option 3:* min(nt, nr)
- *Option 4:* max(nt, nr)

Answer: *Option 3:* min(nt, nr)
Explanation: The maximum number of eigenmodes that the MIMO channel can support is min(nt, nr).
Category: Research publications

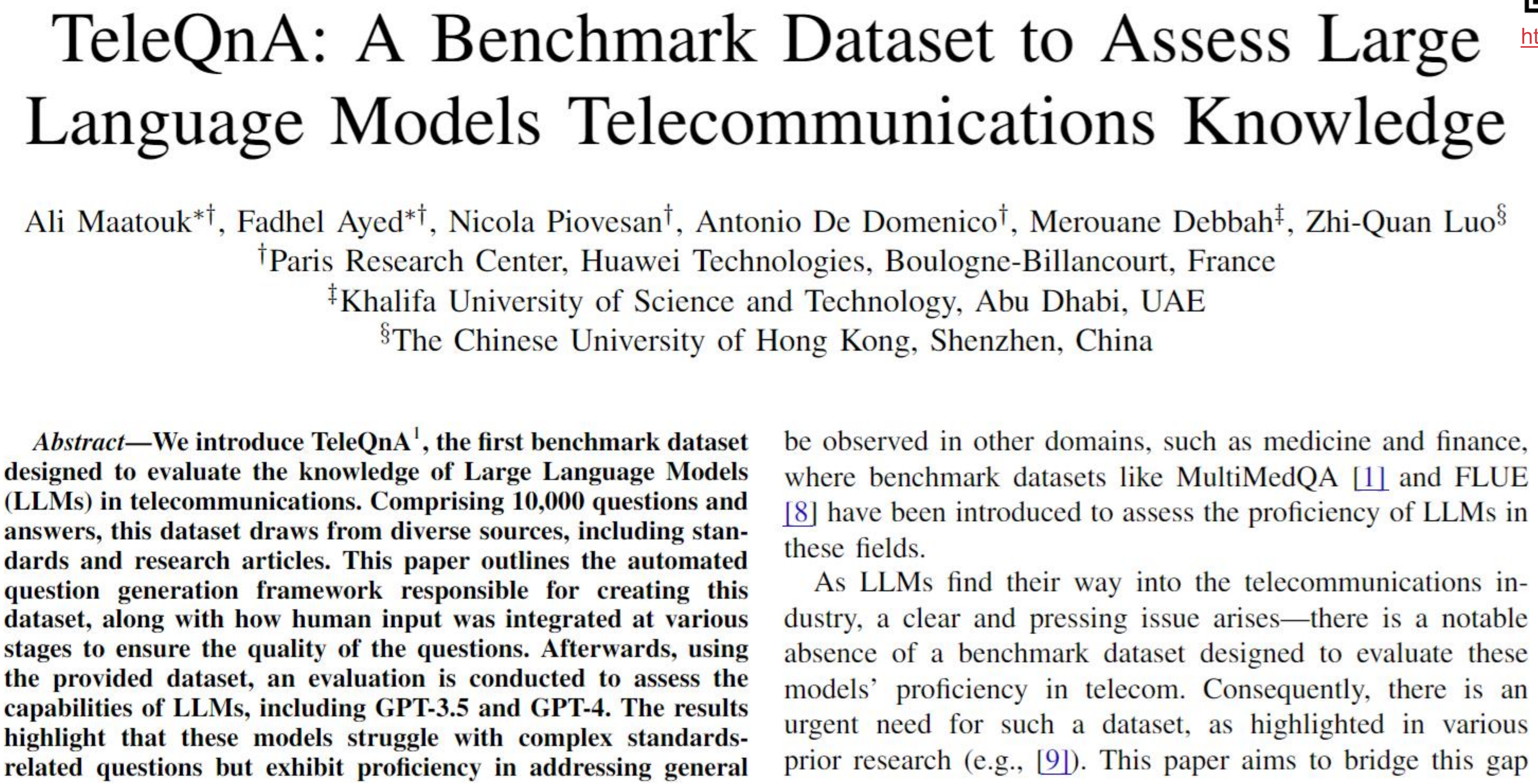
Automatic evaluation based on a LLM-generated dataset with human in the loop!!!

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," in IEEE Network

How can we Evaluate the Telecoms Knowledge of a LLM?

https://huggingface.co/netop

To be published in IEEE Network



TeleQnA: A Benchmark Dataset to Assess Large Language Models Telecommunications Knowledge

Ali Maatouk*[†], Fadhel Ayed*[†], Nicola Piovesan[†], Antonio De Domenico[†], Merouane Debbah[‡], Zhi-Quan Luo[§]
[†]Paris Research Center, Huawei Technologies, Boulogne-Billancourt, France
[‡]Khalifa University of Science and Technology, Abu Dhabi, UAE
[§]The Chinese University of Hong Kong, Shenzhen, China

***Abstract*—We introduce TeleQnA[1], the first benchmark dataset designed to evaluate the knowledge of Large Language Models (LLMs) in telecommunications. Comprising 10,000 questions and answers, this dataset draws from diverse sources, including standards and research articles. This paper outlines the automated question generation framework responsible for creating this dataset, along with how human input was integrated at various stages to ensure the quality of the questions. Afterwards, using the provided dataset, an evaluation is conducted to assess the capabilities of LLMs, including GPT-3.5 and GPT-4. The results highlight that these models struggle with complex standards-related questions but exhibit proficiency in addressing general**

be observed in other domains, such as medicine and finance, where benchmark datasets like MultiMedQA [1] and FLUE [8] have been introduced to assess the proficiency of LLMs in these fields.

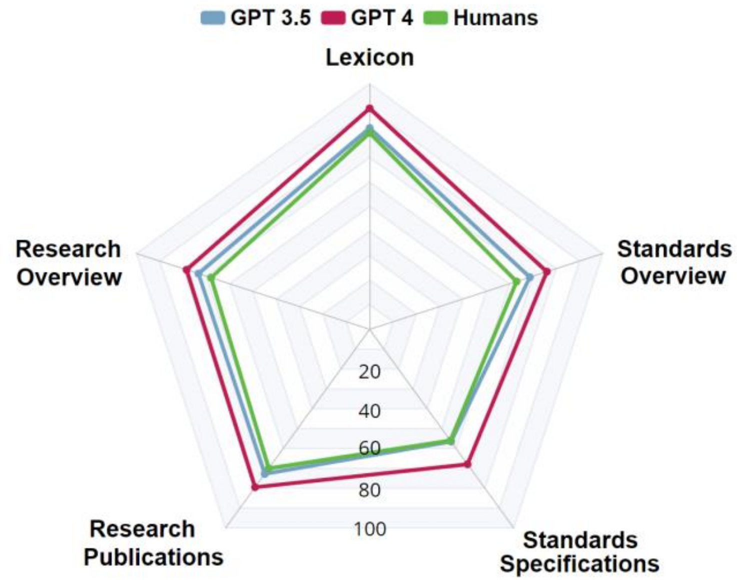
As LLMs find their way into the telecommunications industry, a clear and pressing issue arises—there is a notable absence of a benchmark dataset designed to evaluate these models' proficiency in telecom. Consequently, there is an urgent need for such a dataset, as highlighted in various prior research (e.g., [9]). This paper aims to bridge this gap

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," in IEEE Network

The Telecom Knowledge of GPT

- LLMs exhibit **exceptional performance in the lexicon category**
- LLMs **face challenges when confronted with more intricate questions related to standards**, with the highest performing model, GPT-4, achieving a modest 64% accuracy in this domain
- **LLMs and active professionals exhibit comparable performance in general telecom knowledge**.

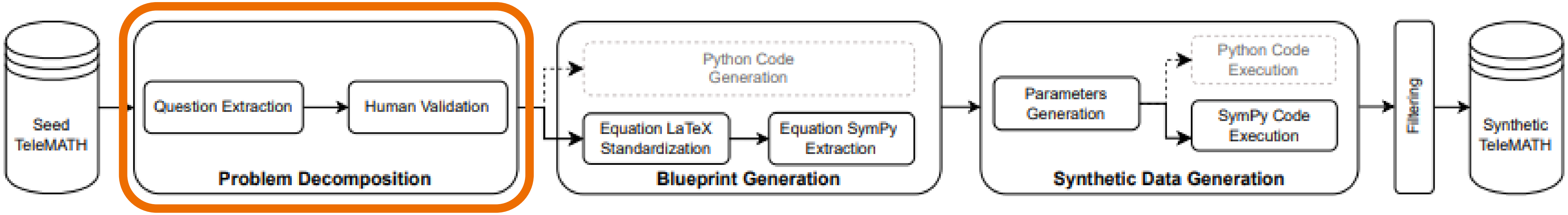
	GPT-3.5	**GPT-4.0**	**Mistral-7B**	**Phi-2**	**Humans**
Lexicon (500)	82.20	**86.80**	69.2	52.60	80.33
Research overview (2000)	68.50	**76.25**	65.9	58.38	63.66
Research publications (4500)	70.42	**77.62**	63.3	54.14	68.33
Standard overview(1000)	64.00	**74.40**	58.8	48.04	61.66
Standard specifications (2000)	56.97	**64.78**	49.7	44.27	56.33
Overall accuracy (10000)	**67.29**	**74.91**	**60.93**	**52.30**	**64.86**

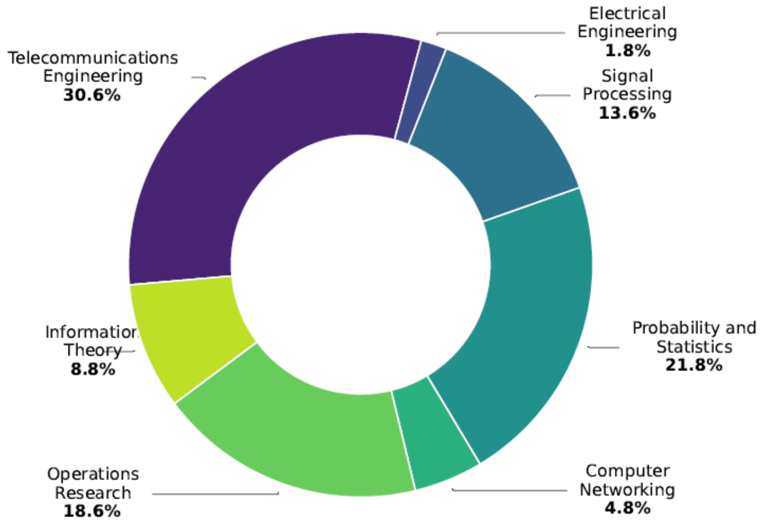


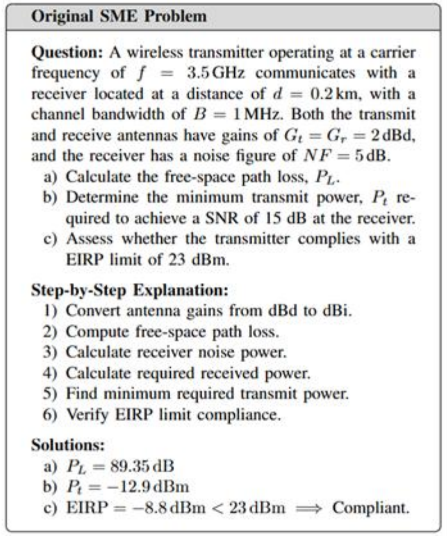
A. Maatouk, F. Ayed, N. Piovesan, A. De Domenico, M. Debbah and Z. -Q. Luo, "TeleQnA: A Benchmark Dataset to Assess Large Language Models Telecommunications Knowledge," in IEEE Network

Benchmarking LLM Capabilities on Telecoms Math

https://huggingface.co/netop







Original SME Problem

Question: A wireless transmitter operating at a carrier frequency of $f = 3.5\,\text{GHz}$ communicates with a receiver located at a distance of $d = 0.2\,\text{km}$, with a channel bandwidth of $B = 1\,\text{MHz}$. Both the transmit and receive antennas have gains of $G_t = G_r = 2\,\text{dBd}$, and the receiver has a noise figure of $NF = 5\,\text{dB}$.

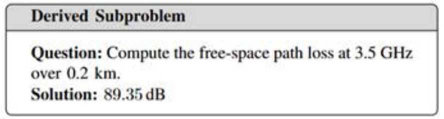
a) Calculate the free-space path loss, P_L.
b) Determine the minimum transmit power, P_t required to achieve a SNR of 15 dB at the receiver.
c) Assess whether the transmitter complies with a EIRP limit of 23 dBm.

Step-by-Step Explanation:

1) Convert antenna gains from dBd to dBi.
2) Compute free-space path loss.
3) Calculate receiver noise power.
4) Calculate required received power.
5) Find minimum required transmit power.
6) Verify EIRP limit compliance.

Solutions:

a) $P_L = 89.35\,\text{dB}$
b) $P_t = -12.9\,\text{dBm}$
c) $\text{EIRP} = -8.8\,\text{dBm} < 23\,\text{dBm} \implies$ Compliant.



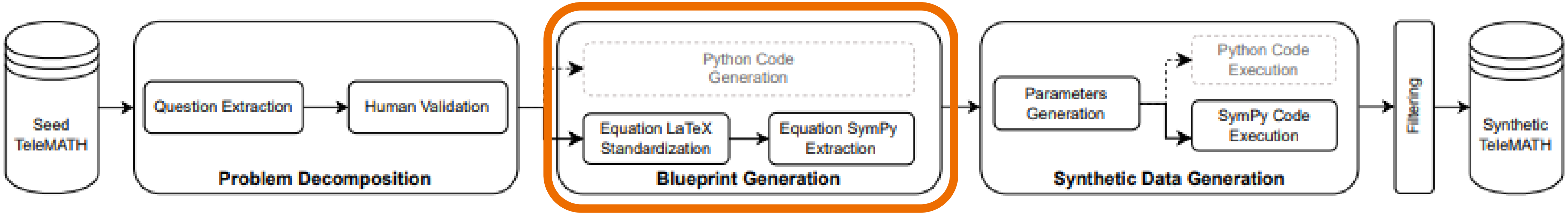
Derived Subproblem

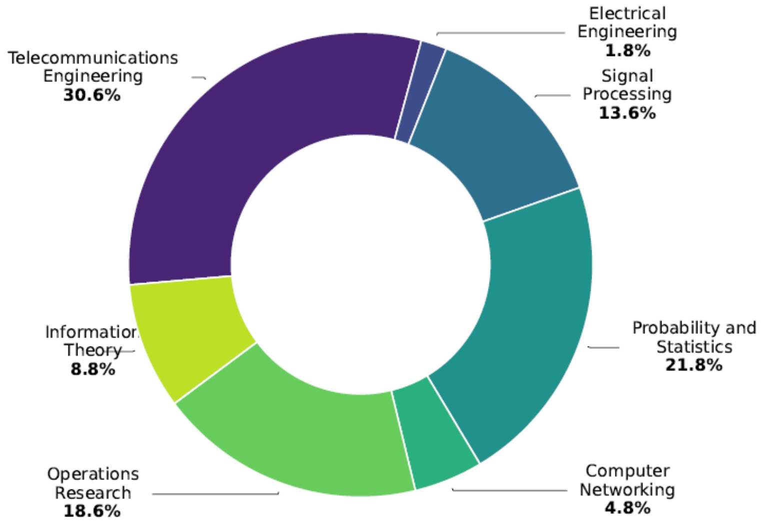
Question: Compute the free-space path loss at 3.5 GHz over 0.2 km.
Solution: 89.35 dB

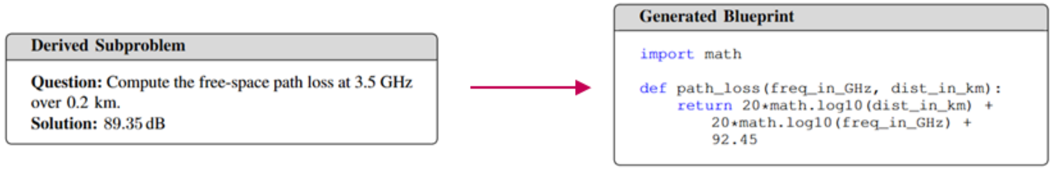
V Colle, M Sana, N Piovesan, A De Domenico, F Ayed, and M. Debbah, "TeleMath: A Benchmark for Large Language Models in Telecom Mathematical Problem Solving," https://arxiv.org/abs/2506.10674

Benchmarking LLM Capabilities on Telecoms Math

https://huggingface.co/netop







Derived Subproblem

Question: Compute the free-space path loss at 3.5 GHz over 0.2 km.
Solution: 89.35 dB

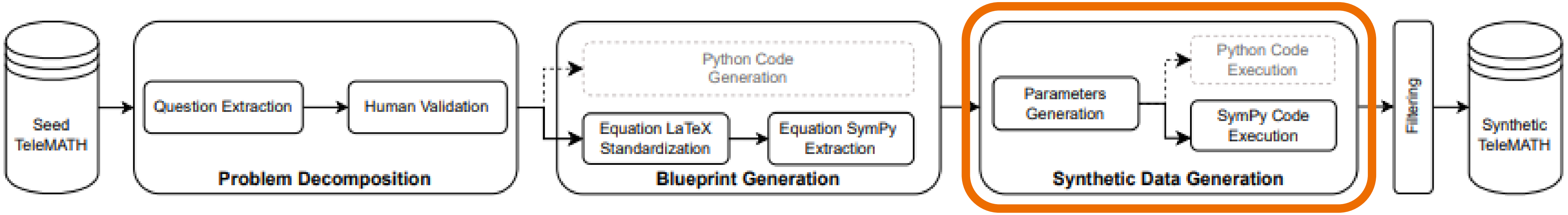
Generated Blueprint

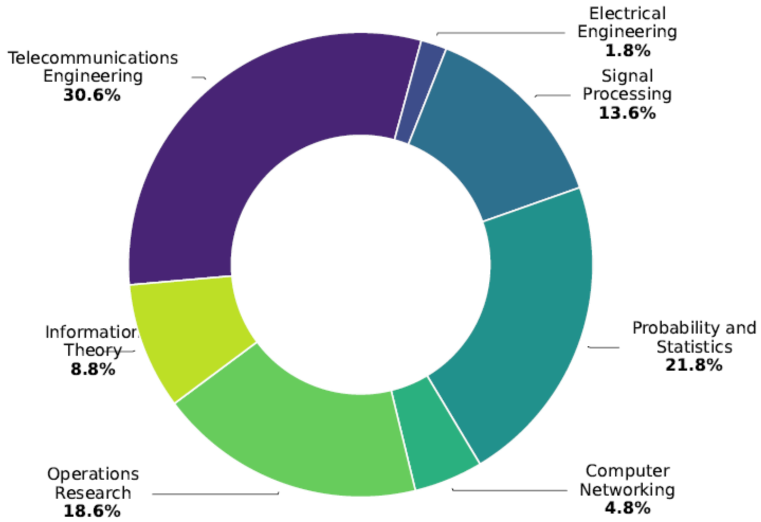
```
import math

def path_loss(freq_in_GHz, dist_in_km):
    return 20*math.log10(dist_in_km) +
        20*math.log10(freq_in_GHz) +
        92.45
```

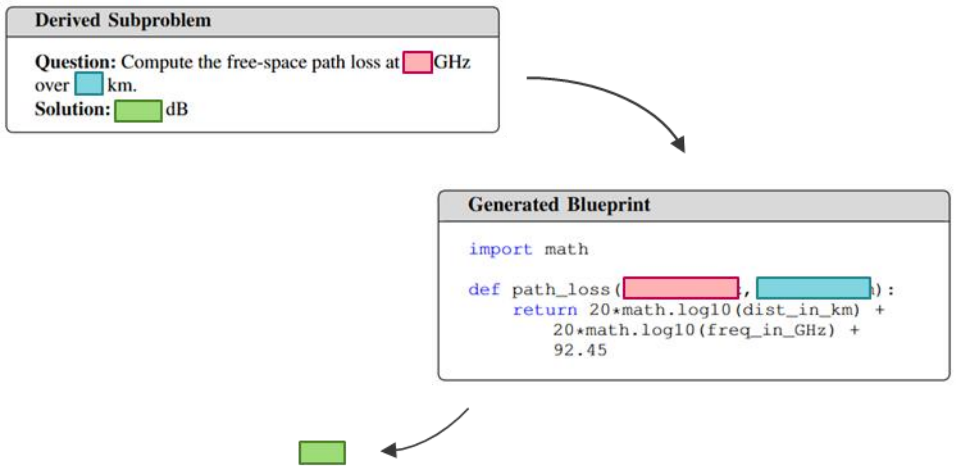
V Colle, M Sana, N Piovesan, A De Domenico, F Ayed, and M. Debbah, "TeleMath: A Benchmark for Large Language Models in Telecom Mathematical Problem Solving," https://arxiv.org/abs/2506.10674

Benchmarking LLM Capabilities on Telecoms Math

https://huggingface.co/netop






Derived Subproblem

Question: Compute the free-space path loss at 🟥 GHz over 🟦 km.
Solution: 🟩 dB

Generated Blueprint

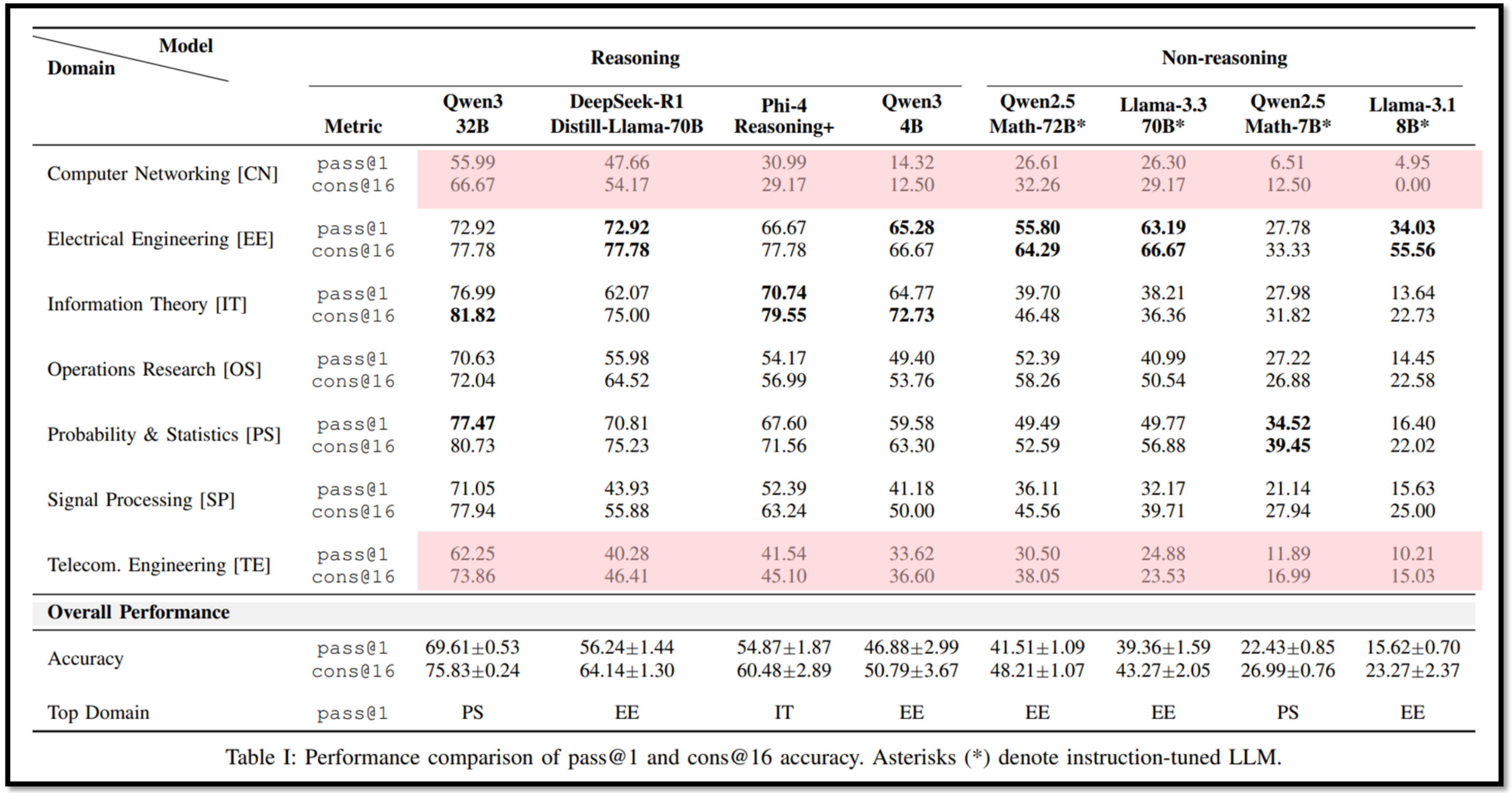
```
import math

def path_loss(🟥, 🟦):
    return 20*math.log10(dist_in_km) +
        20*math.log10(freq_in_GHz) +
        92.45
```

🟩

V Colle, M Sana, N Piovesan, A De Domenico, F Ayed, and M. Debbah, "TeleMath: A Benchmark for Large Language Models in Telecom Mathematical Problem Solving," https://arxiv.org/abs/2506.10674

Benchmarking LLM Capabilities on Telecoms Math

- Reasoning models exhibit **striking performance**
  - But their **inference cost** is much larger that the one of non-reasoning models
- **However, performance are still not acceptable for specific topics:**
  - Computer networking
  - Telecom Engineering



Model / Domain		Reasoning				Non-reasoning			
	Metric	Qwen3 32B	DeepSeek-R1 Distill-Llama-70B	Phi-4 Reasoning+	Qwen3 4B	Qwen2.5 Math-72B*	Llama-3.3 70B*	Qwen2.5 Math-7B*	Llama-3.1 8B*
Computer Networking [CN]	pass@1	55.99	47.66	30.99	14.32	26.61	26.30	6.51	4.95
	cons@16	66.67	54.17	29.17	12.50	32.26	29.17	12.50	0.00
Electrical Engineering [EE]	pass@1	72.92	**72.92**	66.67	**65.28**	**55.80**	**63.19**	27.78	**34.03**
	cons@16	77.78	**77.78**	77.78	66.67	**64.29**	**66.67**	33.33	**55.56**
Information Theory [IT]	pass@1	76.99	62.07	**70.74**	64.77	39.70	38.21	27.98	13.64
	cons@16	**81.82**	75.00	**79.55**	**72.73**	46.48	36.36	31.82	22.73
Operations Research [OS]	pass@1	70.63	55.98	54.17	49.40	52.39	40.99	27.22	14.45
	cons@16	72.04	64.52	56.99	53.76	58.26	50.54	26.88	22.58
Probability & Statistics [PS]	pass@1	**77.47**	70.81	67.60	59.58	49.49	49.77	**34.52**	16.40
	cons@16	80.73	75.23	71.56	63.30	52.59	56.88	**39.45**	22.02
Signal Processing [SP]	pass@1	71.05	43.93	52.39	41.18	36.11	32.17	21.14	15.63
	cons@16	77.94	55.88	63.24	50.00	45.56	39.71	27.94	25.00
Telecom. Engineering [TE]	pass@1	62.25	40.28	41.54	33.62	30.50	24.88	11.89	10.21
	cons@16	73.86	46.41	45.10	36.60	38.05	23.53	16.99	15.03
Overall Performance									
Accuracy	pass@1	69.61±0.53	56.24±1.44	54.87±1.87	46.88±2.99	41.51±1.09	39.36±1.59	22.43±0.85	15.62±0.70
	cons@16	75.83±0.24	64.14±1.30	60.48±2.89	50.79±3.67	48.21±1.07	43.27±2.05	26.99±0.76	23.27±2.37
Top Domain	pass@1	PS	EE	IT	EE	EE	EE	PS	EE

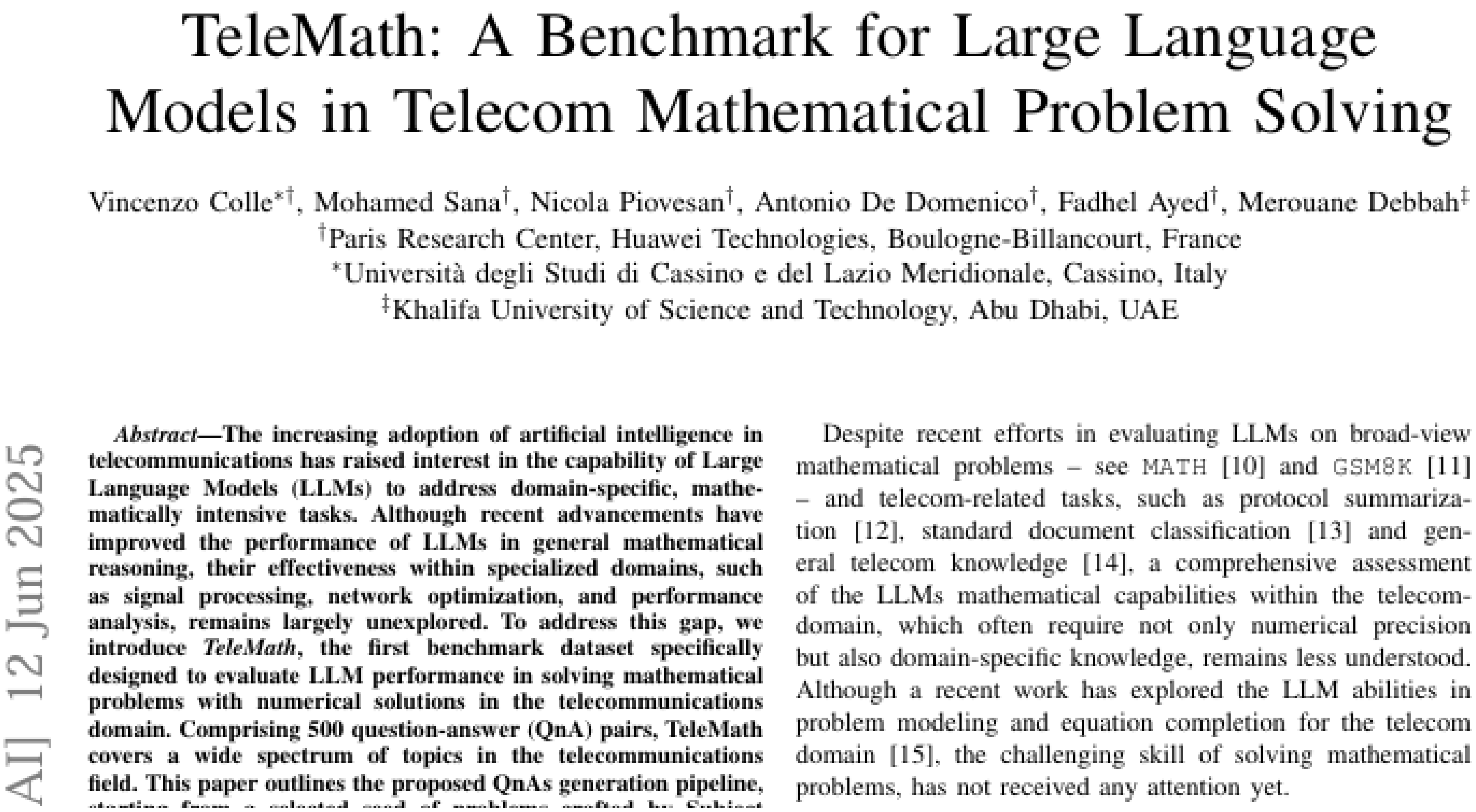
Table I: Performance comparison of pass@1 and cons@16 accuracy. Asterisks (*) denote instruction-tuned LLM.

V Colle, M Sana, N Piovesan, A De Domenico, F Ayed, and M. Debbah, "TeleMath: A Benchmark for Large Language Models in Telecom Mathematical Problem Solving," https://arxiv.org/abs/2506.10674

Benchmarking LLM Capabilities on Telecoms Math

Submitted to IEEE Communications

https://huggingface.co/netop



TeleMath: A Benchmark for Large Language Models in Telecom Mathematical Problem Solving

Vincenzo Colle[*†], Mohamed Sana[†], Nicola Piovesan[†], Antonio De Domenico[†], Fadhel Ayed[†], Merouane Debbah[‡]
[†]Paris Research Center, Huawei Technologies, Boulogne-Billancourt, France
[*]Università degli Studi di Cassino e del Lazio Meridionale, Cassino, Italy
[‡]Khalifa University of Science and Technology, Abu Dhabi, UAE

***Abstract*—The increasing adoption of artificial intelligence in telecommunications has raised interest in the capability of Large Language Models (LLMs) to address domain-specific, mathematically intensive tasks. Although recent advancements have improved the performance of LLMs in general mathematical reasoning, their effectiveness within specialized domains, such as signal processing, network optimization, and performance analysis, remains largely unexplored. To address this gap, we introduce *TeleMath*, the first benchmark dataset specifically designed to evaluate LLM performance in solving mathematical problems with numerical solutions in the telecommunications domain. Comprising 500 question-answer (QnA) pairs, TeleMath covers a wide spectrum of topics in the telecommunications field. This paper outlines the proposed QnAs generation pipeline,**

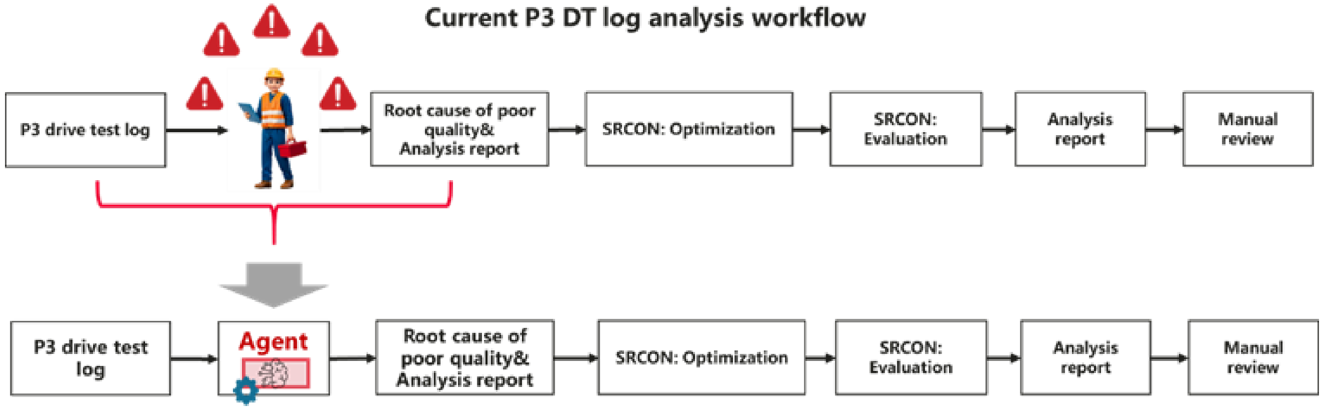
Despite recent efforts in evaluating LLMs on broad-view mathematical problems – see `MATH` [10] and `GSM8K` [11] – and telecom-related tasks, such as protocol summarization [12], standard document classification [13] and general telecom knowledge [14], a comprehensive assessment of the LLMs mathematical capabilities within the telecom-domain, which often require not only numerical precision but also domain-specific knowledge, remains less understood. Although a recent work has explored the LLM abilities in problem modeling and equation completion for the telecom domain [15], the challenging skill of solving mathematical problems, has not received any attention yet.

V Colle, M Sana, N Piovesan, A De Domenico, F Ayed, and M. Debbah, "TeleMath: A Benchmark for Large Language Models in Telecom Mathematical Problem Solving," https://arxiv.org/abs/2506.10674

Can LLMs reduce Costs due to Operations on the Field?

- Several thousands of engineers are engaged in DT data analysis every month, which accounts for nearly 25% of the E2E workload.

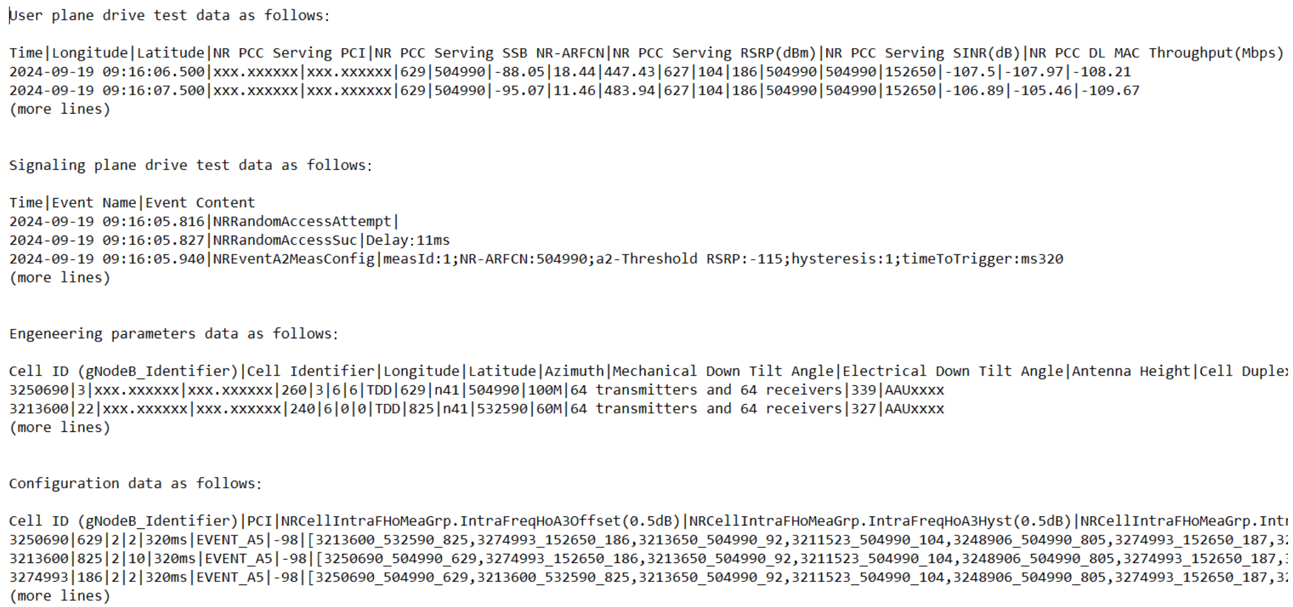
- In addition to the cost, data analysis is complex and lengthy, and its results depend on the engineers' experience and knowledge.



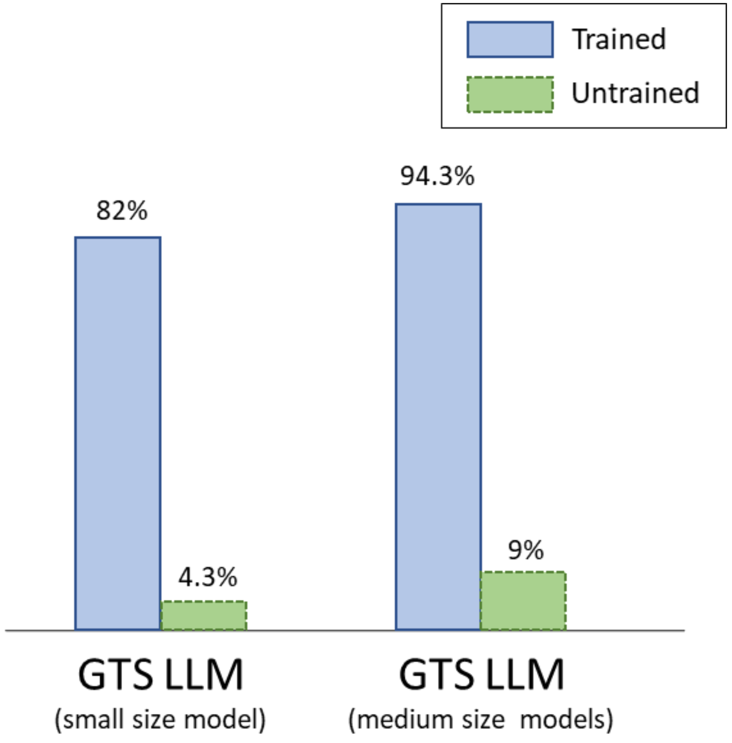
Can LLMs reduce Costs due to Operations on the Field?

Dataset sample:

- Analyze the 5G network data.
- Find the reasons for the low rate (below xxx Mbps) on certain segments.
- Choose the most likely root cause from the following 8 reasons



```
User plane drive test data as follows:

Time|Longitude|Latitude|NR PCC Serving PCI|NR PCC Serving SSB NR-ARFCN|NR PCC Serving RSRP(dBm)|NR PCC Serving SINR(dB)|NR PCC DL MAC Throughput(Mbps)
2024-09-19 09:16:06.500|xxx.xxxxxx|xxx.xxxxxx|629|504990|-88.05|18.44|447.43|627|104|186|504990|504990|152650|-107.5|-107.97|-108.21
2024-09-19 09:16:07.500|xxx.xxxxxx|xxx.xxxxxx|629|504990|-95.07|11.46|483.94|627|104|186|504990|504990|152650|-106.89|-105.46|-109.67
(more lines)


Signaling plane drive test data as follows:

Time|Event Name|Event Content
2024-09-19 09:16:05.816|NRRandomAccessAttempt|
2024-09-19 09:16:05.827|NRRandomAccessSuc|Delay:11ms
2024-09-19 09:16:05.940|NREventA2MeasConfig|measId:1;NR-ARFCN:504990;a2-Threshold RSRP:-115;hysteresis:1;timeToTrigger:ms320
(more lines)


Engeneering parameters data as follows:

Cell ID (gNodeB_Identifier)|Cell Identifier|Longitude|Latitude|Azimuth|Mechanical Down Tilt Angle|Electrical Down Tilt Angle|Antenna Height|Cell Duple
3250690|3|xxx.xxxxxx|xxx.xxxxxx|260|3|6|6|TDD|629|n41|504990|100M|64 transmitters and 64 receivers|339|AAUxxxx
3213600|22|xxx.xxxxxx|xxx.xxxxxx|240|6|0|0|TDD|825|n41|532590|60M|64 transmitters and 64 receivers|327|AAUxxxx
(more lines)


Configuration data as follows:

Cell ID (gNodeB_Identifier)|PCI|NRCellIntraFHoMeaGrp.IntraFreqHoA3Offset(0.5dB)|NRCellIntraFHoMeaGrp.IntraFreqHoA3Hyst(0.5dB)|NRCellIntraFHoMeaGrp.Intr
3250690|629|2|2|320ms|EVENT_A5|-98|[3213600_532590_825,3274993_152650_186,3213650_504990_92,3211523_504990_104,3248906_504990_805,3274993_152650_187,3
3213600|825|2|10|320ms|EVENT_A5|-98|[3250690_504990_629,3274993_152650_186,3213650_504990_92,3211523_504990_104,3248906_504990_805,3274993_152650_187,
3274993|186|2|2|320ms|EVENT_A5|-98|[3250690_504990_629,3213600_532590_825,3213650_504990_92,3211523_504990_104,3248906_504990_805,3274993_152650_187,3
(more lines)
```



Distill (SFT) + RL (HGRPO)

GTS: Global Technical Services

Conclusion

LLMs and agents are expected to have large impact on future networks, both from the requirements and capabilities perspectives

The telcos ecosystem needs to identify the most relevant use cases for LLMs in telcos, and the required functionalities/capabilities proper to telcos (beyond standard NLP)

- Specialized telcos models: Telecom knowledge, coding, reasoning, calculus, etc

Based on these, the industry can define jointly evaluation methodologies:

- Tests, datasets, benchmarking metrics, and platforms
- and creating specialized telcos models

This is just the beginning of the journey

Thank you.

antonio.de.domenico@huawei.com