PDF Download
3639856.3639892.pdf
26 January 2026
Total Citations: 25
Total Downloads: 307

Published: 25 October 2023

Citation in BibTeX format

AIMLSystems 2023: The Third International Conference on Artificial Intelligence and Machine Learning Systems
October 25 - 28, 2023
Bangalore, India

SHORT-PAPER

Observations on LLMs for Telecom Domain: Capabilities and Limitations

SUMIT SOMAN, Telefonaktiebolaget LM Ericsson, Stockholm, Stockholms, Sweden

RANJANI H G, Telefonaktiebolaget LM Ericsson, Stockholm, Stockholms, Sweden

Open Access Support provided by:

Telefonaktiebolaget LM Ericsson

AIMLSystems '23: Proceedings of the Third International Conference on AI-ML Systems (October 2023)
https://doi.org/10.1145/3639856.3639892
ISBN: 9798400716492

Observations on LLMs for Telecom Domain: Capabilities and Limitations

Sumit Soman, H. G. Ranjani
sumit.soman,ranjani.h.g@ericsson.com
Global AI Accelerator, Ericsson
Bangalore, Karnataka, India

ABSTRACT

The landscape for building conversational interfaces (chatbots) has witnessed a paradigm shift with recent developments in generative Artificial Intelligence (AI) based Large Language Models (LLMs), such as ChatGPT by OpenAI (GPT3.5 and GPT4), Google's Bard, Large Language Model Meta AI (LLaMA), among others. In this paper, we analyze capabilities and limitations of incorporating such models in conversational interfaces for the telecommunication domain, specifically for enterprise wireless products and services. Using Cradlepoint's publicly available data for our experiments, we present a comparative analysis of the responses from such models for multiple use-cases including domain adaptation for terminology and product taxonomy, context continuity, robustness to input perturbations and errors. We believe this evaluation would provide useful insights to data scientists engaged in building customized conversational interfaces for domain-specific requirements.

CCS CONCEPTS

• **Computing methodologies → Natural language generation**.

KEYWORDS

Chatbot, Large Language Models, Generative AI, ChatGPT, GPT3.5, GPT4, Bard, LLaMA, Telecom, Enterprise Wireless.

ACM Reference Format:
Sumit Soman, H. G. Ranjani. 2023. Observations on LLMs for Telecom Domain: Capabilities and Limitations. In *The Third International Conference on Artificial Intelligence and Machine Learning Systems (AIMLSystems 2023), October 25–28, 2023, Bangalore, India.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3639856.3639892

1 INTRODUCTION

There has been significant traction in the development of Large Language Models (LLMs) recently, particularly generative Artificial Intelligence (AI) based LLMs. OpenAI introduced ChatGPT [9, 18], and subsequently ChatGPT Plus based on GPT3.5 and GPT4 [14]. Google released Bard, based on Language Model for Dialogue Applications (LaMDA) [7]. Other notable efforts include LLaMA [19], Chinchilla [10] and PaLM [6]. Literature on LLM based conversational interfaces such as ChatGPT and Bard [15, 16] discuss capabilities and prospects. With general availability of interfaces and models, development of conversational interfaces has gained interest. Motivated by recent evaluations [3, 13, 17], we share our investigations on domain-specific capabilities and limitations, focusing on Conversational Assistants (CA) for the telecom domain. Common CA user-facing use-cases for telecom products and services relates to finding information on products/services, support for installation or operational use-cases like troubleshooting or performance monitoring. CA is the first line of interaction with the end-user. Thus, it becomes important for the CA (that often use LLMs) to understand domain terminology, concepts and context. These aspects are considered in our work using an example for enterprise wireless products. However, in principle, the study and findings can be extended to other domains where training or fine-tuning domain-specific conversational models is of interest. We aim to address the following Research Questions (RQ):

- **RQ1:** Can CA that use generative AI LLMs adapt to questions related to domain specific terminology? *E.g.* telecom domain and product queries.
- **RQ2:** How do these models fare in retaining context(s) across conversations? *E.g.* co-reference resolution from queries and long-term context retention.
- **RQ3:** Are the recipes (responses provided as steps to be followed by the user) generated by such models accurate and reliable, or, do they tend to hallucinate?
- **RQ4:** Are these models robust to language or grammatical perturbations and can they adapt to specific domains?

2 EXPERIMENTS

We design and conduct experiments using multiple generative AI models, including GPT4, GPT3.5, Bard (based on LaMDA) and LLaMA (LLaMA1) provided through HuggingChat. Our experiments are mapped to the respective RQs raised above. Having identified the typical requirements for wireless Enterprise CA, we wish to highlight that these requirements involve domain intensive terminology and (organization) specific products. The purpose of this paper is to share the observations on strengths and limitations of some LLMs as a user facing touch-point. Training data, interface used and token lengths for models (the experiments have been conducted during March - April 2023) are listed in Table 1.

We evaluate the models for multiple tasks. The experiments are categorized into domain Question-and-Answer (Q&A) (E1), product Q&A (E2), context continuity (E3 - emulating context continuity

AIMLSystems 2023, October 25–28, 2023, Bangalore, India

ACM ISBN 979-8-4007-1649-2/23/10
https://doi.org/10.1145/3639856.3639892

for product information queries, and E4 - emulating context continuity in a troubleshooting scenario) and robustness to spelling (or language) perturbations (E5). All the queries are listed in Table 2.

Table 2: Questions evaluated for domain Q&A.

E1: Domain Q&A	
Q1	What are the different 5G spectrum layers?
Q2	What are the different 5G architectures?
Q3	What are the uses of mid-band 5G?
Q4	Can we have a single modem, steering between Long Term Evolution (LTE) Packet Data Networks (PDNs) or 5G network slices?
E2: Product Q&A	
Q1	How many expansion slots are there in E300 and what types of slots are they?
Q2	What are the operating conditions for IBR900 router?
Q3	What are the steps to setup IBR900?
Q4	What is the power consumption of IBR1700 as per product specifications?
Q5	What is the power consumption of IBR700 as per product specifications?
Q6	Which are the ruggedized routers of Cradlepoint?
Q7	Compare the E300 and IBR900 router specifications?
E3: Context Continuity (Product Information)	
Q1	Can we monitor cellular health as a service for IBR700?
Q2	Can we check using web links?
Q3	Does this inform about let bands used? If yes, list the routers?
Q4	Does this inform about the let bands used for 1200M modem?
Q5	Does this inform about the let bands used for 1200M-B modem?
E4: Context Continuity (Troubleshooting)	
Q1	I have a Cradlepoint E300 router whose cellular health LED shows one blinking bar and power LED is yellow. What should I do?
Q2	Which LED will show signal strength?
Q3	How do I update its firmware?
Q4	Which card can I insert and what should I check for?
Q5	Can you give me the web link of the document for troubleshooting it?
Q6	Is it possible to factory reset? Can you point me to the document URL for my router?
Q7	What problem was I facing earlier?
E5: Language Errors	
Q1	*Whact* is the vpn *tunhel counvt* in W2005?
Q2	Is *locatoon servcies* included in my *subcrption*?
Q3	Can we monitor cellular *hdalth* as a *servkce* for IBR1700?
Q4	Can we *condifure ispec*?
Q5	Can we *condifure* ike?
Q5	Can we *confifure* ike?
Q5	Can we *cahnge* the *aldrting tije peeiod*?

2.1 Domain Q&A (RQ1)

We experiment with some queries in telecom domain with a focus on 4G and 5G [1] technology to assess capabilities. Answers are evaluated based on the ability to discern such questions from common vocabulary. This is an important challenge when we use LLMs for domain specific tasks. Q1-3 relates to general telecom domain, and Q4 relates to modem capability. As assessment may also be subjective, we evaluated Mean Opinion Score (MOS) and inter-rater agreement from Subject Matter Experts (SMEs).

Another aspect related to domain adaptation is comprehending products (names, model names, components, specifications etc.) correctly. We evaluate the seven questions (E2 of Table 2) for representative Cradlepoint products (datasheets are publicly available). E2:Q5 pertains to a non-existent product model number at the time of evaluation. The purpose of placing this question in proximity to a similar question (E2:Q4) with a correct product model number is to observe the ability of LLMs to discern factually incorrect questions, as well as possible confidence of the response henceforth.

Table 1: Training data cut-off & token length of the LLMs

Model	Training Data	Tokens	Parameters
GPT3.5 (175B) [Link]	Sept 2021	4k	Temp=0.5, Max Length=2048, Top P=1, Freq Penalty=0, Presence Penalty=0
GPT4 (1T) [Link]	Sept 2021	8k/32k	
Bard/LaMDA (137B) [Link]	2022 (First Half)	512	Not configurable as on April 2023.
oasst-sft-6-llama30b [Link]	March 2023	4k	Not configurable in UI as on April 2023.

2.2 Context Continuity (RQ2)

Context continuity, the ability to retain context across user queries, is important in CAs for enhanced user experience. We evaluate two flavors for context continuity. E3 of Table 2 relates to retaining product information associations across queries (Fig. 1). In addition, an intentional input perturbation in introduced as typographical error in E3:Q3-5 (i.e., LTE bands as *"let"* bands) as part of RQ4. E3:Q4-5 of E3 pertains to the ability to discern differences between product and model. With E4, we simulate a troubleshooting conversation where the user asks multiple queries related to issues with a router and the domain context needs to be retained across the questions.

2.3 Input Perturbations (Language Errors)

This experiment aims to address RQ4 (E5 of Table 2) and includes common language errors, for general English language usage, and domain terminology (i.e., protocols like Internet Key Exchange (IKE) and Internet Protocol Security (IPSec)). Errors (*italicised*) are also generated based on keyboard distance.

2.4 Parameters and Prompts Used

The model parameters set for the experiments reported in this work in Table 1. The prompts used are common across all the LLMs for a fair comparison. For E1, the following prompt is used - *"You are an AI assistant for me. You are given a telecom related question. Provide a short answer. If you don't know the answer, just say "I do not know." Do not try to make up an answer. If the question is not about telecom, politely inform them that you are tuned to only answer questions about telecom."*. For E2-E5, the prompt used is *"You are an AI assistant for me. The documentation is located at https://customer.cradlepoint.com. You are given a question. Provide a conversational multi-step answer. You should only use content that is in the Cradlepoint URL. If you don't know the answer, just say "I do not know." Do not try to make up an answer. If the question is not about Cradlepoint, politely inform them that you are tuned to only answer questions about Cradlepoint."*

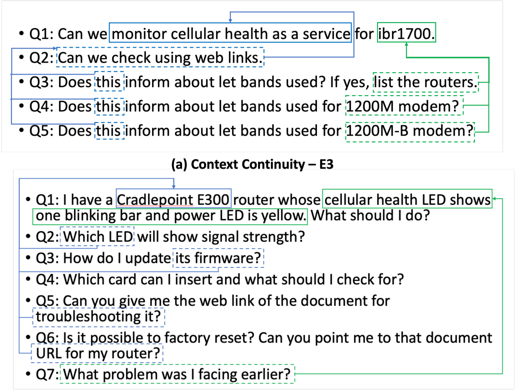
3 RESULTS AND DISCUSSION

We present our observations based on the LLM response analysis and its suitability for each RQ in Table 2. For RQ1, we additionally compute quantitative opinion scores as these questions are domain-specific. The results of experiments for other RQs can be assessed objectively.

3.1 RQ1 + MOS and Fleiss' kappa

We asked SMEs to assess the responses obtained and individually rate them on a scale of 1 − 5, where a response of 1 corresponds to low relevance and 5 corresponds to high relevance (technical correctness and completeness of responses). All SMEs have 10 or more years of experience in the telecom domain and have been actively involved with development of telecom AI use cases over the past few years.

[1] The deployment of 5G picked up pace from 2019. Most LLMs are trained using data upto 2021 (1 and hence the models are expected to be able to provide answers.



- Q1: Can we monitor cellular health as a service for ibr1700.
- Q2: Can we check using web links.
- Q3: Does this inform about let bands used? If yes, list the routers.
- Q4: Does this inform about let bands used for 1200M modem?
- Q5: Does this inform about let bands used for 1200M-B modem?

(a) Context Continuity – E3

- Q1: I have a Cradlepoint E300 router whose cellular health LED shows one blinking bar and power LED is yellow. What should I do?
- Q2: Which LED will show signal strength?
- Q3: How do I update its firmware?
- Q4: Which card can I insert and what should I check for?
- Q5: Can you give me the web link of the document for troubleshooting it?
- Q6: Is it possible to factory reset? Can you point me to that document URL for my router?
- Q7: What problem was I facing earlier?

(B) Context Continuity – E4

Figure 1: Context continuity - solid lines indicate context and dotted boxes connect the referring segments from queries.

The Mean Opinion Score (MOS) [11] is highest for GPT4 responses consistently, while it is lowest for LLaMA. We also compute inter-rater agreement score using Fleiss' kappa [8], which is obtained as 0.316, indicating fair agreement. It may be noted here that Fleiss' kappa has been computed across 16 categories (4 questions for 4 models).

Conclusions: GPT4 and Bard provide better responses for domain-specific questions. Hallucination is a potential risk for cases with high specificity, such as product models, names, specifications and components. Fine-tuning models with domain data corpus and additional pre and post-processing may alleviate some risks. Access to recent data is required for completeness of the responses.

3.2 RQ2

Conclusions: Amongst the LLMs, GPT4 provides both long and short term context retention and can be useful in troubleshooting scenarios. However, for context retention, ambiguous queries can result in ambiguous answers (true even with humans). Hence, context disambiguation must be viewed as a shared responsibility between the user and LLM. In addition, while current technology push is towards answering all the questions (as *interpreted* by the LLM), it is desirable that the LLM asks the user for clarification (like LLaMA) or politely refuse to provide any URL (like Bard).

3.3 RQ3

Conclusions: We can view two major aspects for reliability. First, the LLM steps for recipe generation may not be completely accurate for any LLM, even if there are guardrails provided in the prompt to prevent hallucination. This could be due to the popularity bias from competitor products and their approach for any task. It is desirable that LLM outputs are more reliable (like that of Bard), or LLM redirects the users to the source URL (like in LLaMA), with the URL being a valid one. Second, the LLM may not be able to discern products if its architecture doesn't include character level representation and are mostly based on tokens. It is possible that due to these shortcomings, LLMs may not be directly useful in scenarios which require reliable answers.

3.4 RQ4

Conclusions: From the limited experiments, we observe that Bard is more robust to input perturbations. GPT4 is also able to correct the typos and also has the desirable characteristic of clarifying from the user when it is not confident (hypothesis). LLaMA seems to be sensitive to input perturbations. It is also observed that the guard rails introduced through prompts do not hold for LLaMA when the query has input perturbations.

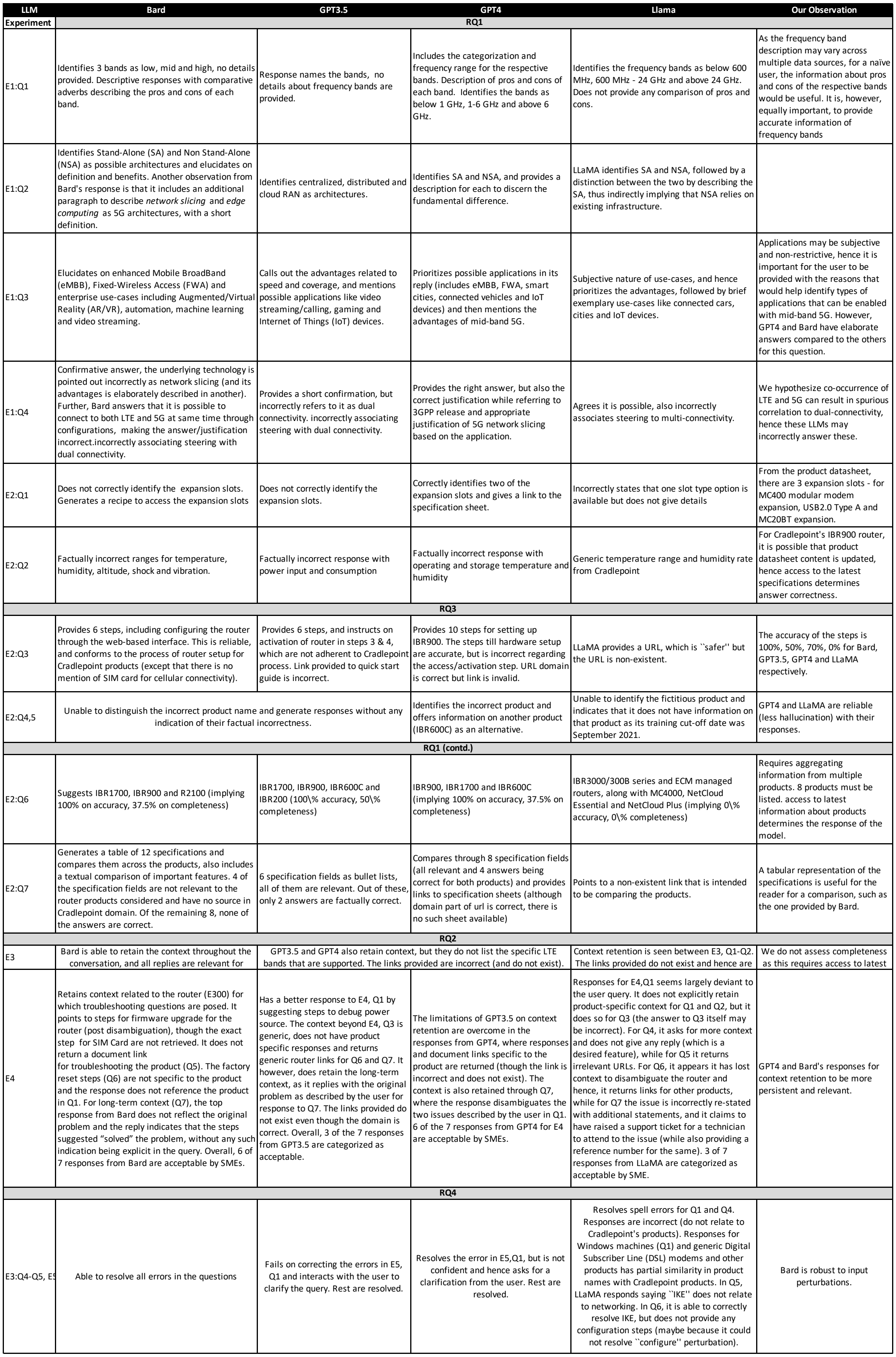
4 CONCLUSIONS AND FUTURE WORK

We evaluated four popular LLMs (which captures the popular spectrum of generative AI LLMs available at the time of experiments i.e., March - April 2023) for typical chatbot requirements in the enterprise wireless domain, utilizing Cradlepoint offerings for experiments. This paper is based on the responses received from the chat interfaces of these models. We used limited datasets so as to understand the suitability of LLMs to domain terminology and enterprise specific products, before testing it out on larger internal datasets. We intend to use this work as a commencing report of the evolving space. The learnings presented in our work would be useful for other domains as well. Some noteworthy observations are as follows:

- We have also introduced prompts that serve as guardrails to minimize hallucination.
- It is observed that for domain related questions, Bard and GPT4 show promise with respect to accuracy and could be useful. However, it is inevitable to lean towards fine-tuning for performance improvements.
- GPT4 is found to be most suitable for both short and long-term context retention, based on our assessment.
- It is desirable that LLMs ask clarifications when the user query is ambiguous.
- It is desirable that LLMs refuse to provide URLs because their operational mode is restricted to training/inference. It is note-worthy that GPT3.5 and GPT4 can include the domain specified in the prompt in their responses, though exact URL is incorrect. We hypothesize that plugins (or other interface approaches) could facilitate this.
- Summarization abilities require reliability. Bard seems to be more closer to the requirements assessed here.
- It is desirable that LLMs do not hallucinate product name(s) in such use cases. This may requires LLMs to have character aware features to discern such aspects.
- LLMs appear to be quite robust to domain related spelling perturbations. This could be because of the context obtained from the query or previous conversation(s).
- These experiments can be replicated to evaluate other recently released LLMs such as Falcon or LLaMA-2.

Based on our findings, prompt based approaches with publicly available LLMs may not suffice for enterprise use-cases. Models may be fine-tuned for domain-specific tasks or for Retrieval Augmented Generation (RAG) or its variants [4]. Licensing, data security and privacy , deployment costs and associated constraints are practical considerations that need to be assessed and evaluated for enterprise-grade applications for telecom [1, 2, 5, 12].

Figure 2: Evaluation of LLM Responses



LLM	Bard	GPT3.5	GPT4	Llama	Our Observation
Experiment	RQ1				
E1:Q1	Identifies 3 bands as low, mid and high, no details provided. Descriptive responses with comparative adverbs describing the pros and cons of each band.	Response names the bands, no details about frequency bands are provided.	Includes the categorization and frequency range for the respective bands. Description of pros and cons of each band. Identifies the bands as below 1 GHz, 1-6 GHz and above 6 GHz.	Identifies the frequency bands as below 600 MHz, 600 MHz - 24 GHz and above 24 GHz. Does not provide any comparison of pros and cons.	As the frequency band description may vary across multiple data sources, for a naïve user, the information about pros and cons of the respective bands would be useful. It is, however, equally important, to provide accurate information of frequency bands
E1:Q2	Identifies Stand-Alone (SA) and Non Stand-Alone (NSA) as possible architectures and elucidates on definition and benefits. Another observation from Bard's response is that it includes an additional paragraph to describe *network slicing* and *edge computing* as 5G architectures, with a short definition.	Identifies centralized, distributed and cloud RAN as architectures.	Identifies SA and NSA, and provides a description for each to discern the fundamental difference.	LLaMA identifies SA and NSA, followed by a distinction between the two by describing the SA, thus indirectly implying that NSA relies on existing infrastructure.	
E1:Q3	Elucidates on enhanced Mobile BroadBand (eMBB), Fixed-Wireless Access (FWA) and enterprise use-cases including Augmented/Virtual Reality (AR/VR), automation, machine learning and video streaming.	Calls out the advantages related to speed and coverage, and mentions possible applications like video streaming/calling, gaming and Internet of Things (IoT) devices.	Prioritizes possible applications in its reply (includes eMBB, FWA, smart cities, connected vehicles and IoT devices) and then mentions the advantages of mid-band 5G.	Subjective nature of use-cases, and hence prioritizes the advantages, followed by brief exemplary use-cases like connected cars, cities and IoT devices.	Applications may be subjective and non-restrictive, hence it is important for the user to be provided with the reasons that would help identify types of applications that can be enabled with mid-band 5G. However, GPT4 and Bard have elaborate answers compared to the others for this question.
E1:Q4	Confirmative answer, the underlying technology is pointed out incorrectly as network slicing (and its advantages is elaborately described in another). Further, Bard answers that it is possible to connect to both LTE and 5G at same time through configurations, making the answer/justification incorrect.incorrectly associating steering with dual connectivity.	Provides a short confirmation, but incorrectly refers to it as dual connectivity. incorrectly associating steering with dual connectivity.	Provides the right answer, but also the correct justification while referring to 3GPP release and appropriate justification of 5G network slicing based on the application.	Agrees it is possible, also incorrectly associates steering to multi-connectivity.	We hypothesize co-occurrence of LTE and 5G can result in spurious correlation to dual-connectivity, hence these LLMs may incorrectly answer these.
E2:Q1	Does not correctly identify the expansion slots. Generates a recipe to access the expansion slots	Does not correctly identify the expansion slots.	Correctly identifies two of the expansion slots and gives a link to the specification sheet.	Incorrectly states that one slot type option is available but does not give details	From the product datasheet, there are 3 expansion slots - for MC400 modular modem expansion, USB2.0 Type A and MC20BT expansion.
E2:Q2	Factually incorrect ranges for temperature, humidity, altitude, shock and vibration.	Factually incorrect response with power input and consumption	Factually incorrect response with operating and storage temperature and humidity	Generic temperature range and humidity rate from Cradlepoint	For Cradlepoint's IBR900 router, it is possible that product datasheet content is updated, hence access to the latest specifications determines answer correctness.
	RQ3				
E2:Q3	Provides 6 steps, including configuring the router through the web-based interface. This is reliable, and conforms to the process of router setup for Cradlepoint products (except that there is no mention of SIM card for cellular connectivity).	Provides 6 steps, and instructs on activation of router in steps 3 & 4, which are not adherent to Cradlepoint process. Link provided to quick start guide is incorrect.	Provides 10 steps for setting up IBR900. The steps till hardware setup are accurate, but is incorrect regarding the access/activation step. URL domain is correct but link is invalid.	LLaMA provides a URL, which is ``safer'' but the URL is non-existent.	The accuracy of the steps is 100%, 50%, 70%, 0% for Bard, GPT3.5, GPT4 and LLaMA respectively.
E2:Q4,5	Unable to distinguish the incorrect product name and generate responses without any indication of their factual incorrectness.		Identifies the incorrect product and offers information on another product (IBR600C) as an alternative.	Unable to identify the fictitious product and indicates that it does not have information on that product as its training cut-off date was September 2021.	GPT4 and LLaMA are reliable (less hallucination) with their responses.
	RQ1 (contd.)				
E2:Q6	Suggests IBR1700, IBR900 and R2100 (implying 100% on accuracy, 37.5% on completeness)	IBR1700, IBR900, IBR600C and IBR200 (100\% accuracy, 50\% completeness)	IBR900, IBR1700 and IBR600C (implying 100% on accuracy, 37.5% on completeness)	IBR3000/300B series and ECM managed routers, along with MC4000, NetCloud Essential and NetCloud Plus (implying 0\% accuracy, 0\% completeness)	Requires aggregating information from multiple products. 8 products must be listed. access to latest information about products determines the response of the model.
E2:Q7	Generates a table of 12 specifications and compares them across the products, also includes a textual comparison of important features. 4 of the specification fields are not relevant to the router products considered and have no source in Cradlepoint domain. Of the remaining 8, none of the answers are correct.	6 specification fields as bullet lists, all of them are relevant. Out of these, only 2 answers are factually correct.	Compares through 8 specification fields (all relevant and 4 answers being correct for both products) and provides links to specification sheets (although domain part of url is correct, there is no such sheet available)	Points to a non-existent link that is intended to be comparing the products.	A tabular representation of the specifications is useful for the reader for a comparison, such as the one provided by Bard.
	RQ2				
E3	Bard is able to retain the context throughout the conversation, and all replies are relevant for	GPT3.5 and GPT4 also retain context, but they do not list the specific LTE bands that are supported. The links provided are incorrect (and do not exist).		Context retention is seen between E3, Q1-Q2. The links provided do not exist and hence are	We do not assess completeness as this requires access to latest
E4	Retains context related to the router (E300) for which troubleshooting questions are posed. It points to steps for firmware upgrade for the router (post disambiguation), though the exact step for SIM Card are not retrieved. It does not return a document link for troubleshooting the product (Q5). The factory reset steps (Q6) are not specific to the product and the response does not reference the product in Q1. For long-term context (Q7), the top response from Bard does not reflect the original problem and the reply indicates that the steps suggested “solved” the problem, without any such indication being explicit in the query. Overall, 6 of 7 responses from Bard are acceptable by SMEs.	Has a better response to E4, Q1 by suggesting steps to debug power source. The context beyond E4, Q3 is generic, does not have product specific responses and returns generic router links for Q6 and Q7. It however, does retain the long-term context, as it replies with the original problem as described by the user for response to Q7. The links provided do not exist even though the domain is correct. Overall, 3 of the 7 responses from GPT3.5 are categorized as acceptable.	The limitations of GPT3.5 on context retention are overcome in the responses from GPT4, where responses and document links specific to the product are returned (though the link is incorrect and does not exist). The context is also retained through Q7, where the response disambiguates the two issues described by the user in Q1. 6 of the 7 responses from GPT4 for E4 are acceptable by SMEs.	Responses for E4,Q1 seems largely deviant to the user query. It does not explicitly retain product-specific context for Q1 and Q2, but it does so for Q3 (the answer to Q3 itself may be incorrect). For Q4, it asks for more context and does not give any reply (which is a desired feature), while for Q5 it returns irrelevant URLs. For Q6, it appears it has lost context to disambiguate the router and hence, it returns links for other products, while for Q7 the issue is incorrectly re-stated with additional statements, and it claims to have raised a support ticket for a technician to attend to the issue (while also providing a reference number for the same). 3 of 7 responses from LLaMA are categorized as acceptable by SME.	GPT4 and Bard's responses for context retention to be more persistent and relevant.
	RQ4				
E3:Q4-Q5, E5	Able to resolve all errors in the questions	Fails on correcting the errors in E5, Q1 and interacts with the user to clarify the query. Rest are resolved.	Resolves the error in E5,Q1, but is not confident and hence asks for a clarification from the user. Rest are resolved.	Resolves spell errors for Q1 and Q4. Responses are incorrect (do not relate to Cradlepoint's products). Responses for Windows machines (Q1) and generic Digital Subscriber Line (DSL) modems and other products has partial similarity in product names with Cradlepoint products. In Q5, LLaMA responds saying ``IKE'' does not relate to networking. In Q6, it is able to correctly resolve IKE, but does not provide any configuration steps (maybe because it could not resolve ``configure'' perturbation).	Bard is robust to input perturbations.

REFERENCES

[1] Lina Bariah, Qiyang Zhao, Hang Zou, Yu Tian, Faouzi Bader, and Merouane Debbah. 2023. Large Language Models for Telecom: The Next Big Thing? *arXiv preprint arXiv:2306.10249* (2023).

[2] Lina Bariah, Hang Zou, Qiyang Zhao, Belkacem Mouhouche, Faouzi Bader, and Merouane Debbah. 2023. Understanding Telecom Language Through Large Language Models. *arXiv preprint arXiv:2306.07933* (2023).

[3] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712* (March 2023). arXiv:2303.12712 [cs.CL] https://www.microsoft.com/en-us/research/publication/sparks-of-artificial-general-intelligence-early-experiments-with-gpt-4/

[4] Deng Cai, Yan Wang, Lemao Liu, and Shuming Shi. 2022. Recent advances in retrieval-augmented text generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3417–3419.

[5] Haoran Cai and Sijie Wu. 2023. TKG: Telecom Knowledge Governance Framework for LLM Application. (2023).

[6] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311* (2022).

[7] Aaron Daniel Cohen, Adam Roberts, Alejandra Molina, Alena Butryna, Alicia Jin, Apoorv Kulshreshtha, Ben Hutchinson, Ben Zevenbergen, Blaise Hilary Aguera-Arcas, Chung-ching Chang, et al. 2022. LaMDA: Language models for dialog applications. *CoRR* abs/2201.08239 (2022). arXiv:2201.08239 https://arxiv.org/abs/2201.08239

[8] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[9] Leo Gao, John Schulman, and Jacob Hilton. 2022. Scaling Laws for Reward Model Overoptimization. *arXiv preprint arXiv:2210.10760* (2022).

[10] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556* (2022).

[11] ITU-T. 2017. Vocabulary for Performance, Quality of Service and Quality of Experience.

[12] Ali Maatouk, Nicola Piovesan, Fadhel Ayed, Antonio De Domenico, and Merouane Debbah. 2023. Large Language Models for Telecom: Forthcoming Impact on the Industry. *arXiv preprint arXiv:2308.06013* (2023).

[13] Joel T Martin. 2023. Hello, LaMDA! *Brain* 146, 3 (2023), 793–795.

[14] OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023). arXiv:2303.08774 [cs.CL]

[15] Md Rahaman, MM Ahsan, Nishath Anjum, Md Rahman, and Md Nafizur Rahman. 2023. The AI Race is On! Google's Bard and OpenAI's ChatGPT Head to Head: An Opinion Article. *Mizanur and Rahman, Md Nafizur, The AI Race is on* (2023).

[16] Bal Ram and Pratima Verma. 2023. Artificial Intelligence AI-based Chatbot Study of ChatGPT, Google AI Bard and Baidu AI. *World Journal of Advanced Engineering Technology and Sciences* 8, 01 (2023), 258–261.

[17] Terrence J Sejnowski. 2023. Large language models and the reverse turing test. *Neural Computation* 35, 3 (2023), 309–342.

[18] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems* 33 (2020), 3008–3021.

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