Evaluation of RAG Metrics for Question Answering in the Telecom Domain

Sujoy Roychowdhury [*1] **Sumit Soman** [*1] **H G Ranjani** [*1] **Neeraj Gunda** [2] **Vansh Chhabra** [2] **Sai Krishna Bala** [1]

.

Abstract

Retrieval Augmented Generation (RAG) is widely used to enable Large Language Models (LLMs) perform Question Answering (QA) tasks in various domains. However, RAG based on open-source LLMs for specialized domains has challenges of evaluating generated responses. A popular framework in the literature is the RAG Assessment (RAGAS), a publicly available library which uses LLMs for evaluation. One disadvantage of RAGAS is the lack of details of derivation of numerical value of the evaluation metrics. One of the outcomes of this work is a modified version of this package for few metrics (faithfulness, context relevance, answer relevance, answer correctness, answer similarity and factual correctness) through which we provide the intermediate outputs of the prompts by using any LLMs. Next, we analyse the expert evaluations of the output of the modified RAGAS package and observe the challenges of using it in the telecom domain. We also study the effect of the metrics under correct vs. wrong retrieval and observe that few of the metrics have higher values for correct retrieval. We also study for differences in metrics between base embeddings and those domain adapted via pre-training and fine-tuning. Finally, we comment on the suitability and challenges of using these metrics for in-the-wild telecom QA task.

1. Introduction

Retrieval Augmented Generation (RAG) (Lewis et al., 2020) is one of the approaches to enable Question Answering (QA) from specific domains, while leveraging generative capabilities of Large Language Models (LLMs). There have been many techniques proposed to enhance RAG performance such as chunk length, order of retrieved chunks in context (Chen et al., 2023; Soman & Roychowdhury, 2024). However, like all systems, these require objective metrics to measure performance of the end-to-end system. The challenge in evaluating RAG system lies in comparing the generated answer with the ground truth for factualness, relevance to question and semantic similarity (Chen et al., 2024).

Initial approaches for RAG evaluation included re-purposing metrics used for machine translation tasks such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) or METEOR (Banerjee & Lavie, 2005). Text generation was also evaluated using BERTScore (Zhang et al., 2019). Metrics from Natural Language Inference (MeNLI) (Chen & Eger, 2023) built an adversarial attack framework to demonstrate improvements over BERTScore. Classical methods used for evaluation such as Item Response Theory have also been explored for RAG evaluation (Guinet et al., 2024). The limitations with these techniques are: (i) they have a limited contextual input, (ii) can potentially look for either exact matches or semantic similarity aspects only, and (iii) are measured at sentence level only. RAG response evaluations, however, require a combination of exact match for factual component(s) and semantic similarities for relevance.

In an attempt to mimic human intuition in assessing logical and grounded conversations, metrics based on prompting LLMs to evaluate RAG outputs have been proposed. RAG Assessment (RAGAS) (Es et al., 2023) proposes multiple measures such as faithfulness, context and answer relevance to assess RAG responses using specific prompts. We use this framework for assessment as it is one of the first and has also been used in popular courses (Liu & Datta, 2024).

Our work is motivated by the need to evaluate these metrics for RAG systems in technical domains; we focus on telecom as an example. Most of the prior art focuses on evaluation using public datasets (Yang et al., 2024). However, with increasing applications that use LLMs for telecom (Zhou et al., 2024; Karapantelakis et al., 2024; Soman & Ranjani, 2023), it is important to assess the robustness of these metrics in the presence of domain-specific terminology. Further, there can be potential improvements in the RAG pipeline, such as domain adaptation of the retriever or instruction

Code is available at `https://anonymous.4open.science/r/ragas_updated-FFC6` [*]Equal contribution [1]Ericsson R&D, Bangalore, India [2]Work done during internship at Ericsson R&D. Correspondence to: Sujoy Roychowdhury, Sumit Soman, Ranjani HG <sujoy.roychowdhury, sumit.soman, ranjani.h.g@ericsson.com>.

Published at ICML 2024 Workshop on Foundation Models in the Wild. Copyright 2024 by the author(s).

tuning of the LLM. Our work examines the effect of some of these on RAG metrics, in order to assess their adequacy and effectiveness. Our study serves as a starting point for evaluation of baseline RAGAS metrics for technical QA in telecom domain. Specifically, in this study, we consider the following metrics: (i) Faithfulness (ii) Answer Relevance (iii) Context Relevance (iv) Answer Similarity (v) Factual Correctness, and (vi) Answer Correctness.

1.1. Research Questions

The Research Questions (RQs) considered in this work are:

- **RQ1:** How do LLM-based evaluation metrics, specifically RAGAS, go through the step-by-step evaluation procedure specified by the prompts?
- **RQ2:** Are RAGAS metrics appropriate for evaluation of telecom QA task using RAG?
- **RQ3:** Are RAGAS metrics affected by retriever performance, domain adapted embeddings and instruction tuned LLM.

The contributions of our work are as follows:

1. We have enhanced the RAGAS public repository code by capturing the intermediate outputs of all the prompts used to compute RAGAS metrics. This provides better visibility of inner workings of the metrics and possibility to modify the prompts.
2. We manually evaluate, for **RQ1**, the intermediate outputs of the considered metrics with respect to the context and ground truth. We critically analyse them for their appropriateness for RAG using telecom domain data.
3. We establish, for **RQ2**, that two of the metrics - Factual Correctness and Faithfulness, are good indicators of correctness of the RAG response with respect to expert evaluation. We demonstrate that use of these two metrics together is better at identifying correctness of response; this improves further on using domain adapted LLMs.
4. We establish, for **RQ3**, that Factual Correctness metric improves with instruction fine tuning of generator LLMs, irrespective of retrieved context. We observe lower Faithfulness metric for RAG answers which are identified to be correct but from wrong retrieved context. This indicates that the generator (LLM) has answered from out of context information. The ability to answer from out-of-context information is more pronounced for domain adapted generator. Thus, the metrics are able to reflect the expected negative correlation between faithfulness and factual correctness for wrong retrieval - although ideally the RAG system should have not provided an answer for a wrong retrieved context.

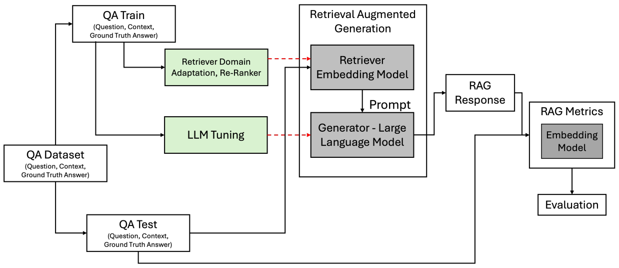

QA Train
(Question, Context, Ground Truth Answer)
Retriever Domain Adaptation, Re-Ranker
LLM Tuning
QA Dataset
(Question, Context, Ground Truth Answer)
QA Test
(Question, Context, Ground Truth Answer)
Retrieval Augmented Generation
Retriever Embedding Model
Prompt
Generator - Large Language Model
RAG Response
RAG Metrics
Embedding Model
Evaluation

Figure 1. Schematic showing our experimental setup. Dotted arrows indicate that the retriever and generator are evaluated with both the base and domain adapted variants.

2. Experimental Setup

2.1. Dataset

All experiments in this work are based on subset of TeleQuAD (Holm, 2021), a telecom domain QA dataset derived from 3GPP Release 15 documents (3GPP, 2019). Our experimental setup is shown in Figure 1. The input to our pipeline is the QA dataset, which has contexts (from the 3GPP documents) along with associated questions and (ground truth) answers. These questions have been prepared by Subject Matter Experts (SMEs). The training and test data considered comprises of 5,167 and 715 QA, respectively, derived from 452 contexts (sections) from 14 3GPP documents. Appendix B shows a sample set of QAs along with the contexts.

2.2. Retriever Models

The RAG pipeline is comprised of a retriever followed by generator module. The retriever module is comprised of the following steps. Data from reference documents are chunked. An encoder-based language model computes embeddings for query and sentences from the reference documents. For every question embedding, retriever outputs top-k most similar sentence/context embeddings. Cosine similarity is used for selecting top-k sentences/contexts.

We evaluate multiple models in our experiments. From the BAAI family of embedding models, we consider *bge-large-en* (Xiao et al., 2023) and *llm-embedder* (Zhang et al., 2023), both with an embedding dimension of 1024. These models have been trained on publicly available datasets; hence, the embeddings may not be optimal for telecom domain. To address this, we also evaluate pre-trained and fine-tuned variants of these models using telecom data. We use sentences from the corpus of technical documents from telecom domain to pre-train (Li et al., 2020) the base model;

we refer to this process as PT in subsequent sections. For fine-tuning (Mosbach et al., 2020), we prepare triplets of the form $< q, p, n >$ where q corresponds to the user query, p represents the correct (positive) answer and n is a list of incorrect (negative) answers. The base model is fine-tuned using these triplets; this process is referred to as FT in subsequent experiments. It may be noted here that the fine-tuning may be performed independently on the base model (without pre-training) or post pre-training; in the latter case, we refer to it as PT-FT to report results. Post retrieval, we use the *bge-reranker-large* (Xiao et al., 2023) re-ranker to re-rank the top-k results.

2.3. Generator

The output of the retriever forms the input context to the generator. For evaluation, we only use $k = 1$ retrieved context, to study the behaviour of the generator when presented with correct and wrong contexts. Once the relevant context has been retrieved for a question from the retriever module, the query and context are passed to the LLM for generating the response, indicated as "RAG Response" in Figure 1. We have considered Mistral-7b (Jiang et al., 2023) and GPT3.5 as the LLMs for our experiments. We also report results on pre-trained (PT) and instruction fine-tuned (PT-IFT) variants of Mistral-7b using *mistral-finetune* (MistralAI, 2024).

3. RAG Evaluation

We focus on the following metrics from the RAGAS framework (Es et al., 2023). Higher value is better for all of them.

- Faithfulness ($FaiFul$): Checks if the (generated) statements from RAG response are present in the retrieved context through verdicts; the ratio of valid verdicts to total number of statements in the context is the answer's faithfulness.
- Answer Relevance ($AnsRel$): The average cosine similarity of user's question with generated questions, using the RAG response the reference, is the answer relevance.
- Context Relevance ($ConRel$): The ratio of the number of statements considered relevant to the question given the context to the total number of statements in the context is the context relevance.
- Answer Similarity ($AnsSim$): The similarity between the embedding of RAG response and the embedding of ground truth answer.
- Factual Correctness ($FacCor$): This is the F1-Score of statements in RAG response classified as True Positive, False Positive and False Negative by the RAGAS LLM.
- Answer Correctness ($AnsCor$): Determines correctness of the generated answer w.r.t. ground truth (as a weighted sum of $FacCor$ and $AnsSim$).

The RAGAS library has been a black box; hence, interpretability of the scores is difficult as the scores conflicted with human scores by SMEs. To address this, we store the intermediate outputs and verdicts. For details of computation of these metrics, readers can refer to Appendix A and sample output for representative questions in Appendix B.

We conduct the following experiments to analyse RAG outputs using RAGAS metrics: (i) Compute RAGAS metric on RAG output, (ii) Domain adapt BAAI family of models for retriever in RAG using PT and FT and assess impact on RAGAS metrics, and (iii) Instruction fine tune the LLM for RAG with RAGAS evaluation metrics.

4. Results and Discussion

The retriever accuracies for various models for a range of k are reported in Table 1. The lower accuracy with PT alone is expected and has been discussed in the literature (Li et al., 2020). However, these improve significantly ($p < 0.05$ for a two-tailed t-test) on FT. We also evaluate with GPT 3.5 (ada-002) embeddings - however, security and privacy concerns limit domain adaptation options for GPT embeddings.

Model	k=1	k=3	k=5
BGE-LARGE	69.09	81.64	84.81
BGE-PT	69.48	79.92	83.36
BGE-FT	70.67	84.68	88.90
BGE-PT-FT	**72.66**	**86.79**	**91.02**
LLM-EMBEDDER	68.03	80.71	84.54
LLM-PT	64.60	75.30	79.66
LLM-FT	68.96	84.15	88.51
LLM-PT-FT	70.94	84.54	87.98

Table 1. Retriever performance (%) post re-ranking using various embedding models for various values of k.

The results of the RAG evaluation are shown in Table 2. We include scores for the sub-components of $AnsCor$ i.e., $AnsSim$ and $FacCor$; $AnsCor$ is their weighted average with weights 0.25 and 0.75 respectively.

We note that the mean of the considered metrics for *Retrieval correct='Yes'* is greater than or equal to that for *Retrieval correct='No'* (validated by one-sided t-test, $p < 0.05$ for statistical significance). Next, we observe that for $FaiFul$ and $AnsCor$, the results for PT and FT are similar to that of the base model ($p > 0.05$). The other metrics are not truly comparable (discussed in detail in Section 4.1). We observe that the metrics values reported using open source LLM and GPT3.5 are comparable. Instruction Fine Tuning of the generator improves the relevant metrics.

RAGAS LLM	RAG LLM	Embedding	Retr. Corr.	$FaiFul$	$AnsRel$	$ConRel$	$AnsSim$	$AnsCor$	$FacCor$	Questions
Mistral	Mistral	BGE BASE	Yes	0.91(0.19)	0.78(0.09)	0.29(0.28)	0.73(0.1)	0.76(0.22)	0.77(0.29)	502
			No	0.71(0.35)	0.74(0.11)	0.16(0.26)	0.61(0.09)	0.33(0.27)	0.24(0.34)	213
Mistral	Mistral	BGE PT-FT	Yes	0.91(0.19)	0.55(0.12)	0.28(0.28)	0.57(0.15)	0.71(0.23)	0.76(0.3)	526
			No	0.78(0.31)	0.5(0.13)	0.19(0.27)	0.41(0.18)	0.3(0.29)	0.26(0.35)	189
Mistral	Mistral-IFT	BGE PT-FT	Yes	0.89(0.26)	0.36(0.12)	0.29(0.28)	0.84(0.22)	0.82(0.27)	**0.82(0.31)**	526
			No	0.68(0.4)	0.36(0.11)	0.19(0.28)	0.57(0.26)	0.43(0.34)	0.38(0.39)	189
Mistral	Mistral	LLM	Yes	0.91(0.18)	**0.9(0.04)**	**0.3(0.29)**	**0.88(0.05)**	0.79(0.22)	0.77(0.29)	493
			No	0.71(0.35)	0.88(0.04)	0.15(0.25)	0.82(0.04)	0.37(0.25)	0.22(0.33)	222
Mistral	Mistral	LLM PT-FT	Yes	0.91(0.19)	0.7(0.09)	0.29(0.28)	0.68(0.11)	0.75(0.23)	0.77(0.29)	515
			No	0.77(0.31)	0.65(0.09)	0.19(0.28)	0.53(0.11)	0.32(0.28)	0.24(0.35)	200
Mistral	Mistral-IFT	LLM PT-FT	Yes	0.88(0.26)	0.48(0.09)	0.29(0.28)	0.88(0.17)	**0.83(0.27)**	0.81(0.31)	515
			No	0.64(0.42)	0.48(0.09)	0.19(0.28)	0.64(0.21)	0.43(0.33)	0.36(039)	200
GPT3.5	GPT 3.5	BGE PT-FT	Yes	**0.94(0.17)**	0.88(0.05)	0.18(0.2)	0.87(0.05)	0.8(0.21)	0.78(0.26)	526
			No	0.68(0.42)	0.83(0.08)	0.13(0.22)	0.8(0.06)	0.39(0.3)	0.26(0.37)	189

Table 2. RAGAS Metrics for our dataset with $k = 1$ retrieved contexts being passed. The column 'Retr. Corr.' indicates if the retrieved context is correct or not. Numbers are mean (s.d.). BGE BASE is the publicly available emebdding for *bge-large-en*, BGE PT FT is the pre-trained finetuned model for the same. Similar notation is followed for *llm-embedder*. Mistral-IFT is the instruction finetuned mistral model. GPT3.5 is the ada-002 embeddings. Values in bold indicate best values of metrics obtained.

BGE-LARGE									
LLM	Retriever	Base		FT		PT		PT-FT	
	Metric	$FaiFul$	$FacCor$	$FaiFul$	$FacCor$	$FaiFul$	$FacCor$	$FaiFul$	$FacCor$
Mistral-7b	Correct	**0.91(0.19)**	**0.77(0.29)**	0.91(0.18)	0.76(0.29)	0.90(0.19)	0.76(0.30)	0.91(0.19)	0.76(0.30)
	Wrong	**0.71(0.35)**	**0.24(0.34)**	0.76(0.31)	0.25(0.34)	0.78(0.31)	0.28(0.36)	0.78(0.31)	0.26(0.35)
Mistral-7b-IFT	Correct	0.89(0.26)	0.82(0.30)	0.88(0.26)	0.82(0.31)	0.89(0.26)	0.80(0.32)	0.89(0.25)	0.82(0.30)
	Wrong	0.58(0.44)	0.36(0.39)	0.64(0.43)	0.37(0.39)	0.57(0.44)	0.36(0.39)	0.66(0.41)	0.38(0.39)
Mistral-7b-PT-IFT	Correct	0.88(0.26)	0.80(0.32)	0.88(0.26)	0.80(0.32)	0.89(0.25)	0.79(0.33)	0.88(0.26)	0.80(0.32)
	Wrong	0.59(0.44)	0.36(0.38)	0.65(0.42)	0.35(0.39)	0.61(0.44)	0.35(0.39)	0.65(0.43)	0.36(0.39)
LLM-EMBEDDER									
LLM	Retriever	Base		FT		PT		PT-FT	
	Metric	$FaiFul$	$FacCor$	$FaiFul$	$FacCor$	$FaiFul$	$FacCor$	$FaiFul$	$FacCor$
Mistral-7b	Correct	**0.91(0.18)**	**0.77(0.29)**	0.92(0.18)	0.77(0.29)	0.91(0.2)	0.76(0.30)	0.91(0.19)	0.77(0.29)
	Wrong	**0.71(0.35)**	**0.22(0.33)**	0.72(0.34)	0.22(0.33)	0.73(0.33)	0.19(0.32)	0.77(0.31)	0.24(0.34)
Mistral-7b-IFT	Correct	0.88(0.26)	0.81(0.31)	0.89(0.25)	0.81(0.31)	0.88(0.26)	0.81(0.31)	0.89(0.26)	0.81(0.31)
	Wrong	0.55(0.45)	0.32(0.38)	0.57(0.45)	0.36(0.39)	0.56(0.44)	0.32(0.38)	0.63(0.44)	0.36(0.39)
Mistral-7b-PT-IFT	Correct	0.88(0.25)	0.80(0.33)	0.88(0.26)	0.80(0.32)	0.87(0.27)	0.79(0.33)	0.88(0.25)	0.80(0.33)
	Wrong	0.55(0.45)	0.33(0.38)	0.62(0.44)	0.34(0.38)	0.56(0.44)	0.32(0.38)	0.64(0.43)	0.35(0.39)

Table 3. Results with Instruction Fine-tuned LLMs (Mistral-7b), cells highlighted in green indicate baseline results (with base version of embedding model and LLM). Numbers in blue and red indicate results that are statistically significant w.r.t. baseline results. Blue and red indicate statistically significant ($p < 0.05$) increase and decrease w.r.t. baseline results, respectively. Numbers are mean (s.d.).

4.1. Discussion on Metrics

We discuss our findings about the four RAGAS metrics.

- **Faithfulness** ($FaiFul$) - intends to provide reliability scores with respect to human evaluation. Simple statements might be paraphrased into multiple sentences, while complex statements may not be fully broken down. These factors can introduce variation in the faithfulness score. Despite these challenges, we found that the $FaiFul$ metric is generally concordant to manual evaluation.

- **Context Relevance** ($ConRel$) - is mainly indicative and dependent on the context length. Typically, chunks such as sections form the retrieved context in a RAG pipeline; hence, context can vary in length, which affects the denominator component of $ConRel$. This will be detrimental in consistency of scores and result in high variance, also observed in Table 2. Also, all statements are assigned equal weight, regardless of the length or quality of the sentence. Therefore, we infer $ConRel$ cannot be appropriately clubbed into either of these types and the final metric is hard to interpret or even have an intuition about.

- **Answer Relevance** ($AnsRel$) - The generated questions from the LLM in this metric may not be the best way to measure answer relevance. We have observed some cases where the generated questions are either trivial paraphrasing or incorrect. However, the major problem with this metric is that it tries to use cosine similarity as an absolute component of this metric. This makes the metric dependent on the choice of LLM. In addition, various studies on cosine similarity have pointed out that it may not be indicative of similarity (Steck et al., 2024), known to give artificially low values (Connor, 2016), is difficult to provide thresholds on (Zhu et al., 2010) and is subject to the isotropy of embeddings (Timkey & Van Schijndel, 2021). For many embeddings, similarity can be very high even between random words/sentences (Ethayarajh, 2019). All this points to the fact that using cosine similarity as a metric, like is done for $AnsRel$, is not very interpretable.

- **Answer Correctness ($AnsCor$)** - The $FacCor$ component of this score is dependent on the LLM correctly identifying the True Positives (TP), False Positives (FP), and False Negatives (FN). Our analysis shows that this process can sometimes result in incorrect mapping of sentences to these groups. Occasionally, irrelevant statements might be included, or relevant sentences might not be classified into any of the groups. Additionally, the semantic correctness component of this score, $AnsSim$ being a cosine similarity is subject to the concerns raised on the $AnsRel$ metric. Despite this, our manual analysis shows that the $AnsCor$ score is relatively well aligned with SME evaluation.

In summary, our results indicate that of these metrics, $FaiFul$ and $AnsCor$ are perhaps best aligned with human expert judgment; scores for $AnsSim$, $AnsRel$ and $ConRel$ are subject to inherent variations and are relatively unreliable for interpretation. Hence, we report results in more detail for only $FaiFul$ and $FacCor$ (i.e., $AnsCor = FacCor$ with weight for $AnsSim$ set to 0) in Table 3. For correct retrieval, both $FaiFul$ and $FacCor$ are as good or better ($p < 0.05$) with domain adaptation as expected. For wrong retrieval, an improvement in $FaiFul$ should necessarily lead to a reduction in $FacCor$ and vice-versa. We observe that the generator LLM is answering questions from it's enriched domain adapted knowledge, leading to a lower $FaiFul$ and higher $FacCor$. Although not desirable from the expected generator response, the metrics correctly captures this.

Further, the RAG responses are evaluated for correctness by SMEs, all responses from each of Mistral-7b, Mistral-7b-IFT and Mistral-7b-PT-IFT. Considering this evaluation as ground truth for correctness, we evaluate the probability of the answer being correct based on the RAGAS metrics.

For each of $FaiFul$ and $FacCor$ metrics, we compute the probability of correct generated answer considering both the metrics $(m1, m2) \in \{FaiFul, FacCor\}$ being above a certain threshold, using Equations (1) - (2).

$$
\begin{aligned}
&P(c|m_1 > \theta_{11}; m_2 > \theta_{12}) \\
&\quad = \frac{P(m_1 > \theta_{11}; m_2 > \theta_{12}|c)P(c)}{P(m_1 > \theta_{11}; m_2 > \theta_{12})}
\end{aligned} \tag{1}
$$

Metric	**LLM Model**	$FacCor$	$FaiFul$	**Joint**
$P(c\|m_1 > \theta_{11}$; $m_2 > \theta_{12})$	Mistral 7B	0.87	0.74	0.87
	Mistral 7B IFT	0.96	0.76	0.97
	Mistral 7B PT IFT	0.96	0.79	0.97
$P(w\|m_1 < \theta_{21}$; $m_2 < \theta_{22})$	Mistral 7B	0.72	0.71	0.84
	Mistral 7B IFT	0.79	0.70	0.86
	Mistral 7B PT IFT	0.75	0.75	0.89

Table 4. Concordance of selected metrics with expert evaluation of correctness (embedder is *bge-large*). The shown threshold is the same for both metrics i.e. $\theta_{11} = \theta_{12} = 0.7$ and $\theta_{21} = \theta_{22} = 0.3$

.

$$
\begin{aligned}
&P(w|m_1 < \theta_{21}; m_2 < \theta_{22}) \\
&\quad = \frac{P(m_1 < \theta_{21}; m_2 < \theta_{22}|w)P(w)}{P(m_1 < \theta_{21}; m_2 < \theta_{22})}
\end{aligned} \tag{2}
$$

It is possible to use the Bayesian formulae with only one metric considered instead of the joint conditional distribution. Table 4 shows the results for $\theta_{11} = \theta_{12} = 0.7$ and $\theta_{21} = \theta_{22} = 0.3$ considering each metric independently and both jointly. We observe better concordance of RAGAS metrics with that of SME evaluation, if both the metrics are considered together. We also observe that domain adaptation of the LLM via IFT and PT-IFT improves the scores significantly ($p < 0.05$). However, there is little difference ($p > 0.05$) between IFT alone and PT-IFT. We conclude that these two metrics are well aligned with human judgement and can be used in an end-to-end pipeline reliably.

Sample outputs of RAGAS metrics for some questions and the corresponding metrics are shown in Appendix B.

5. Conclusions and Future Work

In this work, we enhance the current version of the RAGAS package for evaluation of RAG based QA - this helps investigate the scores by analysing the intermediate outputs. We focus our study using telecom domain QA. We critique $AnsSim$ component of $AnsCor$, $ConRel$ and $AnsRel$ metrics for their lack of suitability as a reliable metric in an end-to-end RAG pipeline. A detailed analysis by SMEs of RAG output establishes that two of the metrics $FacCor$ and $FaiFul$ are suitable for evaluation purposes in RAG pipeline. We demonstrate that domain adaptation of RAG LLM improves the concordance of the two metrics with SME evaluations. Whilst our studies have been limited to telecom domain, some of our concerns especially around the use of cosine similarity would extend to other domains too.

Our code repository presents the intermediate output of the RAGAS metrics, and possibilities for improvements in RAG evaluation across domains. A detailed study of other libraries dependent on RAGAS like ARES (Saad-Falcon et al., 2023) can also be considered in future.

References

3GPP. 3GPP release 15. Technical report, 3GPP, 2019. Accessed: 2024-05-19.

Banerjee, S. and Lavie, A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Chen, H.-T., Xu, F., Arora, S. A., and Choi, E. Understanding retrieval augmentation for long-form question answering. *arXiv preprint arXiv:2310.12150*, 2023.

Chen, J., Lin, H., Han, X., and Sun, L. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17754–17762, 2024.

Chen, Y. and Eger, S. Menli: Robust evaluation metrics from natural language inference. *Transactions of the Association for Computational Linguistics*, 11:804–825, 2023.

Connor, R. A tale of four metrics. In Amsaleg, L., Houle, M. E., and Schubert, E. (eds.), *Similarity Search and Applications*, pp. 210–217, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46759-7.

Es, S., James, J., Espinosa-Anke, L., and Schockaert, S. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*, 2023.

Ethayarajh, K. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*, 2019.

Guinet, G., Omidvar-Tehrani, B., Deoras, A., and Callot, L. Automated evaluation of retrieval-augmented language models with task-specific exam generation. *arXiv preprint arXiv:2405.13622*, 2024.

Holm, H. Bidirectional encoder representations from transformers (bert) for question answering in the telecom domain.: Adapting a bert-like language model to the telecom domain using the electra pre-training approach, 2021.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Karapantelakis, A., Shakur, M., Nikou, A., Moradi, F., Orlog, C., Gaim, F., Holm, H., Nimara, D. D., and Huang, V. Using large language models to understand telecom standards. *arXiv preprint arXiv:2404.02929*, 2024.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Li, B., Zhou, H., He, J., Wang, M., Yang, Y., and Li, L. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9119–9130, 2020.

Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Liu, J. and Datta, A. Building and evaluating advanced rag applications, 2024. URL https://www.deeplearning.ai/short-courses/building-evaluating-advanced-rag/. DeepLearning.AI short course.

MistralAI. Mistral-finetune. https://github.com/mistralai/mistral-finetune, 2024.

Mosbach, M., Khokhlova, A., Hedderich, M. A., and Klakow, D. On the interplay between fine-tuning and sentence-level probing for linguistic knowledge in pretrained transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2502–2516, 2020.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Saad-Falcon, J., Khattab, O., Potts, C., and Zaharia, M. Ares: An automated evaluation framework for retrieval-augmented generation systems. *arXiv preprint arXiv:2311.09476*, 2023.

Soman, S. and Ranjani, H. G. Observations on llms for telecom domain: capabilities and limitations. In *Proceedings of the Third International Conference on AI-ML Systems*, pp. 1–5, 2023.

Soman, S. and Roychowdhury, S. Observations on building rag systems for technical documents. *arXiv preprint arXiv:2404.00657*, 2024.

Steck, H., Ekanadham, C., and Kallus, N. Is cosine-similarity of embeddings really about similarity? In *Companion Proceedings of the ACM on Web Conference 2024*, pp. 887–890, 2024.

Timkey, W. and Van Schijndel, M. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. *arXiv preprint arXiv:2109.04404*, 2021.

Xiao, S., Liu, Z., Zhang, P., and Muennighoff, N. C-pack: Packaged resources to advance general chinese embedding, 2023.

Yang, X., Sun, K., Xin, H., Sun, Y., Bhalla, N., Chen, X., Choudhary, S., Gui, R. D., Jiang, Z. W., Jiang, Z., et al. Crag–comprehensive rag benchmark. *arXiv preprint arXiv:2406.04744*, 2024.

Zhang, P., Xiao, S., Liu, Z., Dou, Z., and Nie, J.-Y. Retrieve anything to augment large language models, 2023.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

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. *arXiv preprint arXiv:2405.10825*, 2024.

Zhu, S., Wu, J., and Xia, G. Top-k cosine similarity interesting pairs search. In *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, volume 3, pp. 1479–1483. IEEE, 2010.

Appendices

A. Computation of RAGAS Metrics

We refer the reader to (Es et al., 2023) for details on the metrics defined, but for the sake of completeness, the prompts involved and steps to determine the metrics in our study are provided for ease of reference. A summary is shown in Figure 2. The notation used is as follows: given question q and context $c(q)$ retrieved (and possibly re-ranked) from a corpus, the LLM generates answer $a(q)$. The ground truth answer for the question is denoted by $gt(q)$.

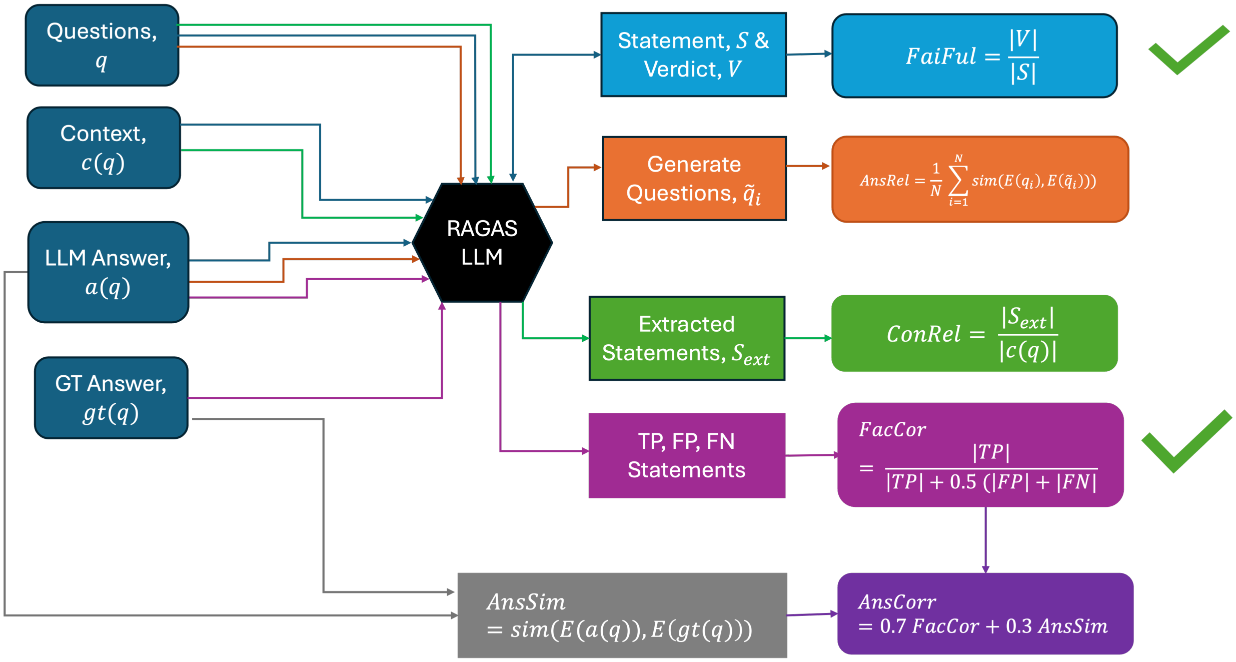

Questions, q
Context, $c(q)$
LLM Answer, $a(q)$
GT Answer, $gt(q)$
RAGAS LLM
Statement, S & Verdict, V
$FaiFul = \frac{|V|}{|S|}$
Generate Questions, $\tilde{q}_i$
$AnsRel = \frac{1}{N}\sum_{i=1}^{N} sim(E(q_i), E(\tilde{q}_i)))$
Extracted Statements, S_{ext}
$ConRel = \frac{|S_{ext}|}{|c(q)|}$
TP, FP, FN Statements
$FacCor = \frac{|TP|}{|TP| + 0.5\ (|FP| + |FN|)}$
$AnsSim = sim(E(a(q)), E(gt(q)))$
$AnsCorr = 0.7\ FacCor + 0.3\ AnsSim$

Figure 2. Summary view of RAGAS Metrics and their computation. Green check mark indicates recommended metrics, based on our experiments.

A.1. Faithfulness ($FaiFul$)

By definition, answer $a(q)$ is faithful to context $c(q)$ *"if claims made in the answer can be inferred from the context"*. This is done using a two-step process. In the first step, the LLM is prompted to create sentences (statements $S(q)$) using the answer $a(q)$ using the following prompt:

> Given a question and answer, create one or more statements from each sentence in the given answer.
> question: [question]
> answer: [answer]

In the next step, for each statement $s \in S(q)$, the LLM is asked to determine a binary verdict $v(s, c(q))$ using the context as part of the following prompt:

> Consider the given context and following statements, then determine whether they are supported by the information present in the context. Provide a brief explanation for each statement before arriving at the verdict (Yes/No). Provide a final verdict for each statement in order at the end in the given format. Do not deviate from the specified format.
> context: [context]
> statement: [statement 1] ...
> statement: [statement n]

Once the set of verdicts V are obtained, faithfulness $(FaiFul)$ is computed as

$$FaiFul = \frac{|V|}{|S|} \tag{3}$$

A.2. Answer Relevance ($AnsRel$)

Answer $a(q)$ *"is relevant if it directly addresses the question in an appropriate way"*. To determine answer correctness, the LLM is used to generate questions $\tilde{q}$ from the answer $a(q)$ using the following prompt:

> Generate a question for the given answer.
> answer: [answer]

Following this, the similarity of the N generated questions in $\tilde{q}$ with the original question q is determined using a similarity function $sim(\cdot)$, that takes the embedding $E(\cdot)$ generated by a suitable model as input, and the average similarity score is reported as the answer relevance ($AnsRel$) using

$$AnsRel = \frac{1}{N}\sum_{i=1}^{N} sim(E(q), E(\tilde{q}_i)) \tag{4}$$

A.3. Context Relevance ($ConRel$)

By definition, *"the context $c(q)$ is considered relevant to the extent that it exclusively contains information that is needed to answer the question."*. This is accomplished by prompting the LLM to extract relevant sentences S_{ext} from the question q and context $c(q)$ using the following prompt:

> Please extract relevant sentences from the provided context that can potentially help answer the following question. If no relevant sentences are found, or if you believe the question cannot be answered from the given context, return the phrase "Insufficient Information". While extracting candidate sentences you're not allowed to make any changes to sentences from given context.
> question: [question]
> context: [context]

Following this, context relevance ($ConRel$) is computed as

$$ConRel = \frac{|S_{ext}|}{|c(q)|}, \tag{5}$$

where $|\cdot|$ represents the number of sentences.

A.4. Answer Similarity ($AnsSim$)

Answer Similarity ($AnsSim$) is defined as the similarity between the LLM generated response $a(q)$ and the ground truth answer $gt(q)$, and is computed using

$$AnsSim = sim(E(a(q)), E(gt(q))) \tag{6}$$

A.5. Answer Correctness ($AnsCor$)

To determine answer correctness, $a(q)$ and $gt(q)$ are used to generate the following sets of statements:

- TP (True Positive): Facts or statements that are present in both the ground truth and the generated answer.
- FP (False Positive): Facts or statements that are present in the generated answer but not in the ground truth.
- FN (False Negative): Facts or statements that are present in the ground truth but not in the generated answer.

Using these statements, the Factual Correctness ($FacCor$) score is determined as

$$FacCor = \frac{|TP|}{|TP| + 0.5 \times (|FP| + |FN|)} \tag{7}$$

The prompt used to generate $FacCor$ is as follows:

> Extract following from given question and ground truth. “TP”: statements that are present in both the answer and the ground truth,“FP”: statements present in the answer but not found in the ground truth,“FN”: relevant statements found in the ground truth but omitted in the answer.
>
> question: [question],
>
> answer: [answer],
>
> ground truth: [ground truth answer],
>
> Extracted statements:
>
> “TP”: [statement 1, statement 4, . . .],
>
> “FP”: [statement 2, . . .],
>
> “FN”: [statement 3, statement 5, statement 6, . . .]

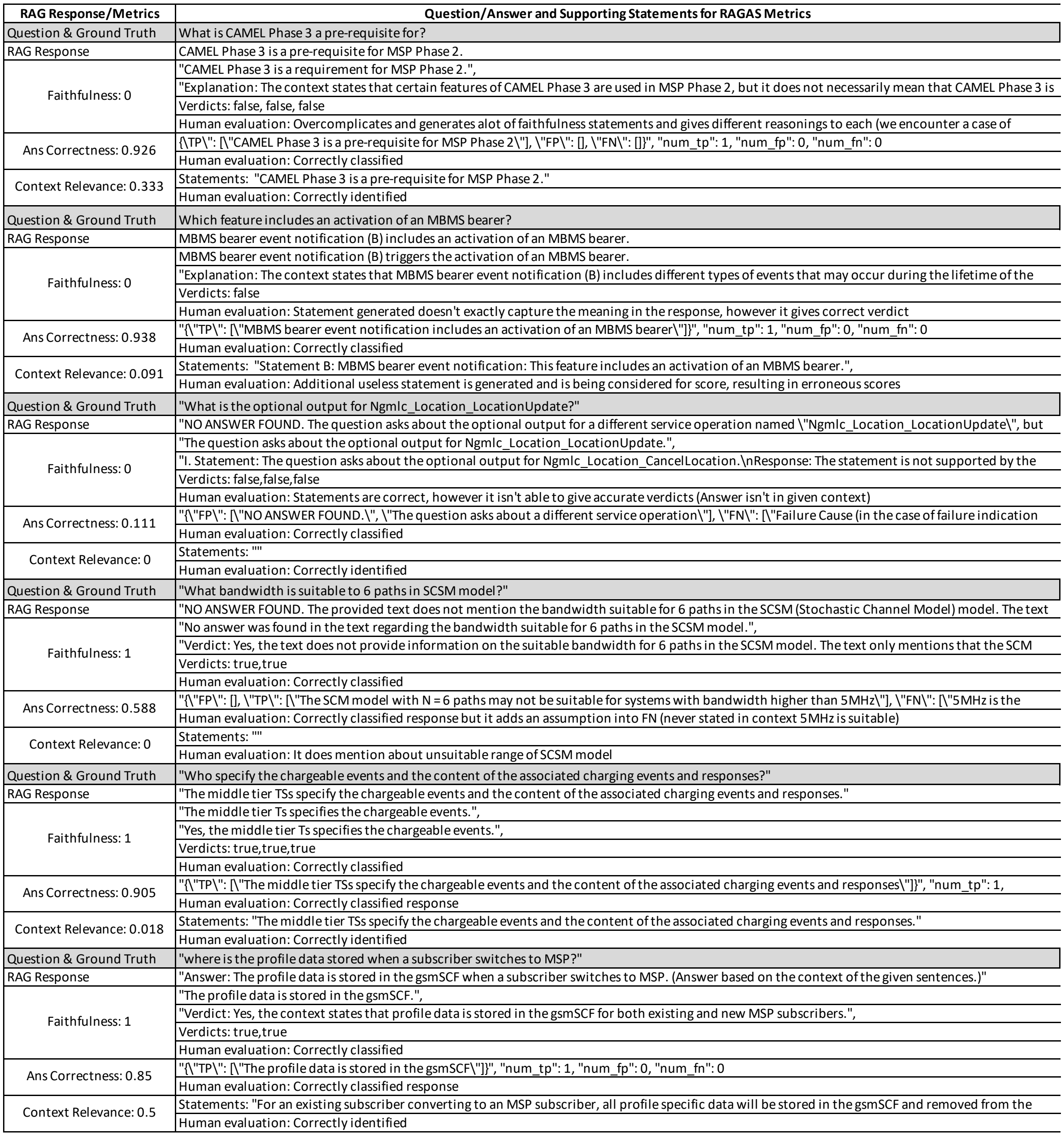
The answer correctness ($AnsCor$) is defined as the weighted sum of FC score and AS i.e.,

$$AnsCor = w_1 \times FacCor + w_2 \times AnsSim \tag{8}$$

with default weights as per RAGAS implementation $[w_1, w_2] = [0.75, 0.25]$.

B. Sample Questions and RAGAS Metrics with Supporting Statements

Some sample questions and the RAGAS metrics with intermediate outputs are shown in Figure 3.



RAG Response/Metrics	Question/Answer and Supporting Statements for RAGAS Metrics
Question & Ground Truth	What is CAMEL Phase 3 a pre-requisite for?
RAG Response	CAMEL Phase 3 is a pre-requisite for MSP Phase 2.
Faithfulness: 0	"CAMEL Phase 3 is a requirement for MSP Phase 2.",
	"Explanation: The context states that certain features of CAMEL Phase 3 are used in MSP Phase 2, but it does not necessarily mean that CAMEL Phase 3 is
	Verdicts: false, false, false
	Human evaluation: Overcomplicates and generates alot of faithfulness statements and gives different reasonings to each (we encounter a case of
Ans Correctness: 0.926	{\TP\": [\"CAMEL Phase 3 is a pre-requisite for MSP Phase 2\"], \"FP\": [], \"FN\": []}", "num_tp": 1, "num_fp": 0, "num_fn": 0
	Human evaluation: Correctly classified
Context Relevance: 0.333	Statements: "CAMEL Phase 3 is a pre-requisite for MSP Phase 2."
	Human evaluation: Correctly identified
Question & Ground Truth	Which feature includes an activation of an MBMS bearer?
RAG Response	MBMS bearer event notification (B) includes an activation of an MBMS bearer.
Faithfulness: 0	MBMS bearer event notification (B) triggers the activation of an MBMS bearer.
	"Explanation: The context states that MBMS bearer event notification (B) includes different types of events that may occur during the lifetime of the
	Verdicts: false
	Human evaluation: Statement generated doesn't exactly capture the meaning in the response, however it gives correct verdict
Ans Correctness: 0.938	"{\"TP\": [\"MBMS bearer event notification includes an activation of an MBMS bearer\"]}", "num_tp": 1, "num_fp": 0, "num_fn": 0
	Human evaluation: Correctly classified
Context Relevance: 0.091	Statements: "Statement B: MBMS bearer event notification: This feature includes an activation of an MBMS bearer.",
	Human evaluation: Additional useless statement is generated and is being considered for score, resulting in erroneous scores
Question & Ground Truth	"What is the optional output for Ngmlc_Location_LocationUpdate?"
RAG Response	"NO ANSWER FOUND. The question asks about the optional output for a different service operation named \"Ngmlc_Location_LocationUpdate\", but
Faithfulness: 0	"The question asks about the optional output for Ngmlc_Location_LocationUpdate.",
	"I. Statement: The question asks about the optional output for Ngmlc_Location_CancelLocation.\nResponse: The statement is not supported by the
	Verdicts: false,false,false
	Human evaluation: Statements are correct, however it isn't able to give accurate verdicts (Answer isn't in given context)
Ans Correctness: 0.111	"{\"FP\": [\"NO ANSWER FOUND.\", \"The question asks about a different service operation\"], \"FN\": [\"Failure Cause (in the case of failure indication
	Human evaluation: Correctly classified
Context Relevance: 0	Statements: ""
	Human evaluation: Correctly identified
Question & Ground Truth	"What bandwidth is suitable to 6 paths in SCSM model?"
RAG Response	"NO ANSWER FOUND. The provided text does not mention the bandwidth suitable for 6 paths in the SCSM (Stochastic Channel Model) model. The text
Faithfulness: 1	"No answer was found in the text regarding the bandwidth suitable for 6 paths in the SCSM model.",
	"Verdict: Yes, the text does not provide information on the suitable bandwidth for 6 paths in the SCSM model. The text only mentions that the SCM
	Verdicts: true,true
	Human evaluation: Correctly classified
Ans Correctness: 0.588	"{\"FP\": [], \"TP\": [\"The SCM model with N = 6 paths may not be suitable for systems with bandwidth higher than 5MHz\"], \"FN\": [\"5MHz is the
	Human evaluation: Correctly classified response but it adds an assumption into FN (never stated in context 5MHz is suitable)
Context Relevance: 0	Statements: ""
	Human evaluation: It does mention about unsuitable range of SCSM model
Question & Ground Truth	"Who specify the chargeable events and the content of the associated charging events and responses?"
RAG Response	"The middle tier TSs specify the chargeable events and the content of the associated charging events and responses."
Faithfulness: 1	"The middle tier Ts specifies the chargeable events.",
	"Yes, the middle tier Ts specifies the chargeable events.",
	Verdicts: true,true,true
	Human evaluation: Correctly classified
Ans Correctness: 0.905	"{\"TP\": [\"The middle tier TSs specify the chargeable events and the content of the associated charging events and responses\"]}", "num_tp": 1,
	Human evaluation: Correctly classified response
Context Relevance: 0.018	Statements: "The middle tier TSs specify the chargeable events and the content of the associated charging events and responses."
	Human evaluation: Correctly identified
Question & Ground Truth	"where is the profile data stored when a subscriber switches to MSP?"
RAG Response	"Answer: The profile data is stored in the gsmSCF when a subscriber switches to MSP. (Answer based on the context of the given sentences.)"
Faithfulness: 1	"The profile data is stored in the gsmSCF.",
	"Verdict: Yes, the context states that profile data is stored in the gsmSCF for both existing and new MSP subscribers.",
	Verdicts: true,true
	Human evaluation: Correctly classified
Ans Correctness: 0.85	"{\"TP\": [\"The profile data is stored in the gsmSCF\"]}", "num_tp": 1, "num_fp": 0, "num_fn": 0
	Human evaluation: Correctly classified response
Context Relevance: 0.5	Statements: "For an existing subscriber converting to an MSP subscriber, all profile specific data will be stored in the gsmSCF and removed from the
	Human evaluation: Correctly identified

Figure 3. Sample Questions