**3GPP TSG-SA WG1 Meeting #108**  **S1-244513**

**Orlando, USA, 18-22 November 2024** *(revision of S1-244082)*

Title: System knowledge as part of Retrieval Augmented Generation for Generative AI

Agenda Item: 8.1.7

Source: Nokia

Contact: Hideaki Takahashi; hideaki dot takahashi at nokia dot com

*Abstract: This use case proposes to leverage system knowledge of telecommunications as part of Retrieval Augmented Generation for Generative AI.*

Updates in S1-244513:

- Add an example scenario how 6G system knowledge can be leveraged for RAG.

---------- Use Case template ----------

FIRST CHANGE

## 3.3 Abbreviations

For the purposes of the present document, the abbreviations given in 3GPP TR 21.905 [1] and the following apply. An abbreviation defined in the present document takes precedence over the definition of the same abbreviation, if any, in 3GPP TR 21.905 [1].

LLM Large Language Model

LMM Large Multimodal Model

RAG Retrieval-Augmented Generation

SLM Small Language Model

SECOND CHANGE (NEW TEXT)

## Y.x Use case on System knowledge as part of Retrieval Augmented Generation for Generative AI

### Y.x.1 Description

Generative AI is an approach that aims at creating a new content of different kinds mimicking the characteristics of the training data. One of the types of Generative AI is Large Language Model (LLM), which is based on the natural language/text and is used to generate plausible language as output based on the user’s query. Some examples of such models are different versions of GPT, Llama, etc. The generic LLM, also called foundation model, is usually obtained by exhaustive training using enormous amount of data. This process is regarded as pre-training. In order to adapt such generic LLM to specific task or domain, such as understanding technical text and recommending management actions in telco domains, the fine-tuning of foundation model needs to be performed. This can be done by training the model on task and domain specific data, selectively training a subset of model parameters or a set of newly added parameters, etc.

LLMs use language as input and output data modality. However, this is not the only modality of data and approach under umbrella of Generative AI. For example, Large Multimodal Models (LMM) combine various data modalities, e.g. text, audio, visual, etc. capturing the correlations between different data modalities. Small Language Models (SLM) are less compute intense than LLMs, both in training and inference, but may still achieve satisfying performance especially if trained and applied for specific problem, etc.

Retrieval-Augmented Generation (RAG) is an approach for improving the quality of LLM-generated outputs by grounding the model on the knowledge sources external to the model itself. In other words, RAG approach enables access to the information beyond training data, thus supplementing the model’s internal representation of information. It consists of two phases, cf. Figure Y.x.1-1:

1) Retrieval phase: In this phase the search and retrieval of information snippets of most relevance to the user’s prompt is done. This may be performed by different algorithms e.g., similarity scoring with cosine calculation of user input and external information. The retrieved external information is appended to the user’s prompt and given to the model.

2) Generation phase: In this phase the LLM leverages on its generative capabilities as well as the augmented prompt during the retrieval phase in order to provide the final output to user prompt.

RAG ensures that the LLM has access to the most relevant and up-to-date facts important for output generation, thereby improving the quality of the output. In addition, the RAG approach lowers the costs, including the energy consumption of updating the LLM model (e.g., with respect to re-training/fine-tuning using large amount of data being performed continuously, periodically or over extensive time windows).

For example, 6G system can leverage on RAG approach in order to support augmentation of the user prompts towards Generative AI models supporting XR service for city sightseeing. Alice is in the train on her way from Minich to Paris, where she will spend a weekend in sightseeing the city. Alice asks site sightseeing application powered by Generative AI model to provide her with the summary of different insights on the city during her train trip. She requests insights on history, geographical, population aspects as well as the recommendations on city highlights to visit, including the XR preview of recommendations (e.g. museum and gallery visits, concert or sport events visits). In such way Alice can have a glimpse into the experience that she might have when actually visiting the recommended highlights and decide, already while being in the train, in interaction with the XR service, what to visit during the weekend in Paris. In order to ensure realistic and seamless glimpse on expected experience the XR service may involve generating and fetching data of different modalities and potentially large size. The XR service can leverage on the RAG and available knowledge sources from MNO’s network to get the information on the network conditions in order to better plan fetching the content needed for interactive content generation and seamless and cost-conscious delivery to Alice. Such information may include, e.g. roaming conditions and agreements in different areas on the way from Munich to Paris that may impact the way Alice will be charged, coverage map, known coverage issues in different areas om the way, e.g. due to geographical circumstances, high mountain peaks, tunnels, availability of different technologies, on the way such as 3G, 4G, 5G, etc. Based on this information retrieved from available knowledge sources from MNO’s network, XR service can enable seamless interactive experience for Alice, even while traversing different areas of different network conditions.



Figure Y.x.1-1. Retrieval Augmented Generation for LLM

**Potential sustainability impacts of the use case**

|  |  |  |  |
| --- | --- | --- | --- |
|  | (the UN SDGs/GDC matching goals of each aspect within 3GPP context) | **Potential benefits of the use case (added value)** | **Potential areas of attention of the use case (risks to be mitigated)** |
| **Environmental sustainability aspects**(UN SDGs 12, 13, 14, 15 and indirectly 6, 7 & 11UN GDC “Develop principles for environmental sustainability of digital technologies”) | **Energy resources**(UN SDG 7, 11, 12) | * Reducing energy consumption by enabling high quality and up-to-date model outputs without need of energy demanding re-training or fine-tuning, but by providing access to the most relevant and up-to-date system knowledge.
 | * Potentially higher energy consumption to realize the retrieval augmented generation using system knowledge sources.
 |
| **Emissions**(UN SDG 6, 7, 11, 12, 13, 14, 15) | * Enabling CO2 emission reduction due to the benefit of reduced energy consumption.
 | * Potential CO2 emission as a result of realization of retrieval augmented generation using system knowledge sources
 |
| **Socio-economic sustainability aspects**(UN SDGs 2, 3, 4, 5, 8, 9, 10, 11, 16 & 17 and indirectly 12)UN GDC “Closing Digital Divides and Accelerating SDG Progress”&“Expanding Digital Economy Inclusion”&“Fostering an Inclusive, Safe Digital Space”) | **Inclusion & Equality** **(**UN SDGs 11, 10, 4, 5 and indirectly 3, 16 & 17) | * Allows to better take the user context (e.g. system knowledge) into account by augmenting GenAI prompts, thus better adapting to its conditions.
 |  |
| **Trustworthiness (**UN SDGs 11 and indirectly 3 & 17) | * Increasing trustworthiness by grounding the models on the most reliable and up-to-date knowledge sources for generating the outputs.
 | Risks of knowledge source misuse or breach, e.g., by unauthorized consumers. |

### Y.x.2 Pre-conditions

An MNO or an application service provider (ASP) deploys one or more system knowledge sources at various locations, which are used as part of “RAG” to augment the user prompts towards Generative AI models supporting different services, e.g., consumer XR services, vertical industry services such as production lines or site inspection etc.

Subject to MNO’s or ASP’s policy, system knowledge sources are always in-operation and offering different system knowledge for augmenting the Generative AI prompts related to different services when requested by the users. Such system knowledge sources can be of different types, containing different network data, e.g. from different network domains, etc, located in different areas and owned by different parties, e.g. MNO or application service providers. Furthermore, the system knowledge sources may be associated with certain restrictions or recommendations for usage as well as the cost, e.g. in terms of delay in providing the generated output or monetary cost of using the system knowledge source.

A user subscribes to application services (e.g., Generative AI based XR applications, or vertical industry services etc.) provided by the MNO and/or the ASP. The ASP has an agreement with the MNO to leverage MNO’s network information as RAG knowledge source(s), which are being exposed by the MNO. For instance, Alice subscribes an XR service for city sightseeing powered Generative AI models from MNO and/or ASP. Alice is in the train on her way from Minich to Paris, where she will spend a weekend in sightseeing the city.

### Y.x.3 Service Flows

1. The user invokes the subscribed application and requests the Generative AI based service to MNO’s or ASP’s application host. The user prompts the Generative AI model, e.g. LLM as part of the invoked application. E.g. Alice invoked the XR application for city sightseeing. She requests via user prompts to Generative AI insights on history, geographical, population aspects as well as the recommendations on city highlights to visit, including the XR preview of recommendations (e.g. museum and gallery visits, concert or sport events visits).

2. Upon the user prompt, the Generative AI model, e.g. an LLM supporting the application, invokes the RAG (Retrieval Augmented Generation) approach to access the MNO’s network information as knowledge source in order to provide most up-to-date and reliable output (e.g. answer or action) to user prompt.

3. MNO’s network exposes information about the available system knowledge sources that can be used as part of RAG. Such information may include description of the system knowledge sources, type of knowledge contained, including different network data, from different network domains, address or area where the knowledge is stored, entity owning the knowledge source, description on usage restrictions, associated cost, etc. For instance, the exposed information includes roaming conditions and agreements in different areas on the way from Munich to Paris that may impact the way Alice will be charged, coverage map, known coverage issues in different areas om the way, e.g. due to geographical circumstances, high mountain peaks, tunnels, availability of different technologies, on the way such as 3G, 4G, 5G, etc.

4. The application chooses to use the available knowledge sources from MNO’s network or not to use them based on the exposed information, e.g. constraints and costs.

5. If the application chooses to use the available knowledge sources from MNO’s network, the application selects the desired system knowledge source to be used for augmenting the output generation by the Generative AI model, e.g. LLM supporting the application. E.g. the XR application from Alice, selects the knowledge source containing the MNO’s network information related to e.g. roaming conditions and agreements, coverage map, known coverage issues, availability of different technologies etc. In addition, the application can select how and to what extent the system knowledge source should be used, e.g. in terms of the target and maximum amount of data that should be retrieved for generating the output.

6. The MNO’s network enables the required measurements related to knowledge source usage by the application, further monitoring if the available knowledge sources were used for augmenting the output generation and deriving the related statistics

7. The MNO’s network provides the reports on conducted measurements and statistics related to usage of available knowledge sources towards the ASP, further charging the ASP on a pay-per-use basis. The ASP may further use the reported measurements relating them with other indications such as output quality, delay in generating the output, cost etc. Based on such insights on relation between system knowledge source usage and the output quality or costs the ASP may adjust its choice of available knowledge sources from MNO’s network .

### Y.x.4 Post-conditions

The user can enjoy optimal user experience as the Generative AI output (e.g. answers or actions) as part of the application were taken based on most up-to-date and reliable knowledge sources from the network. E.g. Alice enjoys realistic and seamless insights into the Paris highlights already while being on the way to Paris, and decides what to visit based on interests, time constraints, budget, etc.

In this example, the outputs from Alice’s XR application may contain the actions on content transmissions which counteracts up-to-date information on expected network issues or unfavourable roaming conditions. For some other examples, such as for vertical industry applications the output may contain the fact of pausing or taking out of production lines some machines that need maintenance, initiate software updates if needed, based on up-to-date troubleshooting tickets resolution records etc.

In addition, the MNO and ASP can contribute to save energy consumption by means of the on-demand knowledge augmentation of Generative AI supporting the user application, without the need for energy consuming processes, such as re-training and fine-tuning of the models.

### Y.x.5 Existing features partly or fully covering the use case functionality

TS 22.261 has defined requirements and KPIs for AI/ML model transfer, but not for RAG use cases.

### Y.x.6 Potential New Requirements needed to support the use case

[PR Y.x.6-1] Subject to regulatory requirements, operator’s policy and user consent, the 6G system shall be able to expose towards authorized 3rd parties the capability to support RAG for GenAI services through the usage of 6G system knowledge.

NOTE: Exposed 6G system knowledge can contain, e.g. type of knowledge (e.g. from different network domains, such as RAN, Core, OAM, subscription data, etc), entity owning the knowledge, access or usage restrictions, associated latency or cost, etc.

[PR Y.x.6-2] Subject to operator’s policy, the 6G system shall enable an authorized consumer to discover and access 6G system knowledge capability and sources (e.g. in terms of type or description of available knowledge, such as relating to different network domains, address or area where knowledge is stored, entity owning the knowledge, etc.)as needed for GenAI prompt augmentation.

[PR Y.x.6-3] Subject to operator’s policy, the 6G system shall enable an authorized consumer to request how and to what extent 6G system knowledge should be used for GenAI prompt augmentation.

[PR Y.x.6-4] Subject to operator’s policy, the 6G system shall enable network operators to configure which 6G system knowledge is authorized to which 3rd party for GenAI prompt and output augmentation.

[PR Y.x.6-5] Subject to operator’s policy, the 6G system shall enable measurements and statistics related to 6G system knowledge usage.

[PR Y.x.6-6] Subject to regulatory requirements and operator’s policy, the 6G system shall be able to expose reports to an authorized 3rd party on the measurements and statistics related to this 3rd party’s 6G system knowledge usage for GenAI prompt and output augmentation.

[PR Y.x.6-7] Subject to regulatory requirements and operator’s policy, the 6G system shall be able to charge an authorized 3rd party based on its 6G system knowledge usage for GenAI prompt and output augmentation.