**3GPP TSG-SA WG4 Meeting #130S4-241952**

**Orlando, Florida, USA, 18 - 22 November 2024**

**Source: Samsung Electronics Co., Ltd, Interdigital New York, Huawei**

**Title: [FS\_AI4Media] pCR on conclusions**

**Agenda item: 9.6**

**Document for: Agreement**

**1. Introduction**

PD v1.4.0 contains some initial conclusions for TR 26.927. This contribution provides a draft text to be discussed and included into TR 26.927.

**2. Discussion**

The text describes the work which has been studied and addressed by the TR, possible areas for further study, and possible next steps.

**3. Proposal**

It is proposed to agree the following changes to 3GPP TR 26.927 v0.9.0.

\* \* \* First Change \* \* \*

## 10 Conclusions

AI/ML in media services involve the use of AI/ML models to perform media processing, typically with video or audio media as the input into an AI/ML model, giving an output which may be processed video or audio media (or even a different media type), or a specific description of the input media itself, such as labelling in object recognition. In order to support such AI/ML based media processing, UE devices may support on device AI inferencing, but depending on UE AI capabilities, support for AI inferencing in the network may also be required for use cases where on device AI inferencing may be difficult or infeasible.

In this study, the broad findings for AI/ML model transfer in TR 22.874 [aa] have been further analysed with specific focus on media-based AI/ML use cases and scenarios, in particular considering how AI/ML models and data may be distributed over the 5G system, the feasibility and implications of splitting AI/ML operations between different AI/ML endpoints (noticeably the UE and the network), and the compression of AI/ML model data and intermediate data. Due to the broad range of applications for AI/ML based media processing, as well as the wide diversity of different AI/ML models available for each same application, feasibly evaluations for a given set of scenarios are also included in TR 26.847 [xx] as part of this study.

Based on the core use cases, basic functional architectures are presented for basic AI/ML model distribution, split AI/ML operation and distributed/federated learning, with the introduction of the different AI user plane data components involved (noticeably AI model data, intermediate data, inference input and output data), and the definition of a set of logical AI functions.

The identified logical AI functions are further mapped to the 5G system, addressing the underlying 5GMS/RTC and IMS DC architectures. The mapping of such AI media use cases to the different architectures and their relevant procedures describe the need for the provisioning, capability discovery/negotiation and delivery session support for the delivery of AI data components and the use of required AI media functions at different endpoints according to the service configuration negotiated. Architecture variants for three different collaboration scenarios are also introduced, each with a different level of MNO network support for AI/ML functions.

[Based on the details in the report, the following is identified:

Possible work in the short-term:

* For collaboration scenarios 1 (Over The top) and 2 (Hosting), document the relevant procedures to support the delivery of AI data components based on the architectures in TS 26.501 and TS 26.506 for 5GMS and RTC respectively
* For collaboration scenario 3 (MNO-operated), document the relevant procedures to support the configuration and operation of split AI inferencing between the UE and the network, considering UE on-device AI capability, according to the feasible use cases and scenarios identified, based on the architectures in TS 26.501/TS 26.506 and TS 23.228 for 5GMS/RTC and IMS DC respectively
* Define one or more 3GPP interoperable formats for the AI data components associated with the relevant AI media services
* Investigate and identify the need to extend existing APIs at referencing points to support the delivery of AI data components
* Further identify the necessary interoperable metadata to enable the configuration, delivery and processing of AI data components by different endpoints, namely between the UE and the network, based on the initial findings in this document
* Further investigate mechanisms to deliver the required metadata according to the associated architectures used for the AI media service, including the use of existing interfaces and reference points

Possible work in the mid-term:

* Further investigate and study the impacts and needs for the compression of certain AI data components
* Further evaluate state-of-the-art AI models and their impacts on the requirements of existing use cases and scenarios defined in this document and in TR 26.847 [xx]
* Evaluate any new use case and scenarios relevant to collaboration scenario 3, including distributed/federated learning]

\* \* \* End of Changes \* \* \*