**3GPP TSG-SA WG4 Meeting #131S4-250112**

**Geneva, Switzerland, 17 - 21 February 2025**

**Source: Samsung Electronics Co., Ltd, Tencent**

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

**Agenda item: 9.6**

**Document for: Agreement**

**1. Introduction**

This contribution provides a slight update to the conclusions in TR 26.927 for discussion.

**2. Discussion**

The group should discuss and agree on a way forward for the conclusion, which is relevant to the timeline of future work and studies for AI in SA4.

**3. Proposal**

It is proposed to agree the following changes to 3GPP TR 26.927 v0.10.0.

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

# 8 Conclusion

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 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:

- For collaboration scenarios 1 (Over The top) and 2 (Hosting):

- document the traffic characteristics of the identified AI/ML media data for the relevant use cases.

- identify any potential needs for new QoS identifier and/or QoS procedures to support the delivery of the identified AI/ML data based on the architectures in TS 26.501, TS 26.506, and TS 26.114 for 5GMS, RTC, and IMS respectively

- For collaboration scenario 3 (MNO-operated):

- identify, e.g. based on the recent SA2 studies and IMS telephony evolution, the need for MNO-operated AI/ML services

- document and potentially extend the relevant procedures to support the configuration and operation of these services.

- If need for split inference in these services is identified, enhance procedures to support 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

For the two bullet points above:

- Whenever necessary, reference or specify interoperable formats for the AI data

-

- Whenever necessary, identify the necessary interoperable metadata to enable the configuration, delivery and processing of AI data

- Further investigate mechanisms to deliver the required metadata according to the associated architectures

[Other possible work:

- Further investigate and study the impacts and needs for the compression of certain AI data including models.

- 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 cases and scenarios relevant to collaboration scenario 3, including distributed/federated learning

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