**3GPP TSG-SA WG4 Meeting #131S4-250112\_r02**

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

**Source: Samsung Electronics Co., Ltd, Tencent, Interdigital, Qualcomm Inc. HHI, Nokia Denmark, V[odafone Telekomünikasyon A.S.](https://portal.etsi.org/webapp/TelDir/QueryOrgaInfo.asp?OrgaId=12486" \t "_blank)**

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

**Agenda item: 9.6**

**Document for: Agreement**

**1. Introduction**

This contribution provides a merge between the S4-250112. Rev2 and S4-250328.Rev2

**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 the latest version of the functional PD.

\* \* \* 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, resulting in an output which may be a version of processed video or audio media (e.g. picture enhancement, audio translation), a specific description of the input media itself (e.g. labelling in object recognition) or a completely new media (e.g. sign/text translation to speech or video). In order to support such AI/ML based media processing, three scenarios have been documented:

* UE device AI inferencing
* AI inferencing in the network
* Split AI inferencing between the UE and the network.

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 this document describes how AI/ML models and data may be distributed over the 5G system, and documents the split 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 same application, feasibly evaluations for a given set of scenarios are documented in TR 26.847 [xx] as part of this study.

Based on the core use cases, functional architectures are presented for basic AI/ML model distribution, split AI/ML operation and distributed/federated learning. Different AI user plane data components have been identified and documented (AI model data, intermediate data, inference input and output data), and a set of logical AI functions have been defined.

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 next steps are identified:

[Normative in release-20]:

1. For collaboration scenario 3 IMS services:

- Recommend stage 3 normative work on the support of AI/ML model distribution and operation in IMS.

- Extend TS 26.114 and TS 26.264 specifications to support AI/ML data delivery and AI/ML media processing in IMS services, as identified in clauses 5.4 and 5.5 respectively.

- Specify support for AI/ML data signalling and negotiation, including support for split inferencing based on the split configurations and model compositions as described in clause 5.1.1.

- Select interoperable formats for AI model data and intermediate data as described in clauses 6.2 and 6.3.

* Specify the mechanism to address the necessary and different interoperable compression profiles to enable the configuration, delivery and processing of AI data. NOTE: details of this bullet point needs further discussion before moving from the PD to the TR.

[Beyond release-19]:

1. For collaboration scenarios 1 (OTT) and 2 (Hosting):

- Further study, identify and document the traffic characteristics of the AI/ML data components (as defined in clause 5.3.1 and detailed in clause 6) for the relevant use cases, as introduced in TR 26.847.

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

1. For collaboration scenario 3 non-IMS services:

- Further study and investigate stage 2 aspects for the architectures in TS 26.501 (5GMS) and TS 26.506 (RTC), identifying potential key issues related to:

1) The support of AI model distribution and operation, based on details in clause 5.3.6.

2) The support of split AI inferencing between the UE and the network, based on details in clause 5.3.5.

3) The support of distributed/federated learning, in particular SA2 defined features, as identified in clause 5.3.7.

1. Other possible work:
* 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

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