**3GPP TSG-SA WG4 Meeting #131S4-250238\_Rev1**

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

**Source: Interdigital Finland Oy, Fraunhofer HHI, Nokia Denmark**

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

**Agenda item: 9.x**

**Document for: Agreement**

**1. Introduction**

This contribution provides an update to the conclusions section of TR 26.927 and proposes to:

* Remove the brackets to define a normative phase for the procedures supporting AI/ML model and split inferencing data delivery for media services.
* Delete the point regarding investigating the impact and the needs for compression of certain AI data components, as TR 26.847 has demonstrated its feasibility.
* Add a new point identifying the need to specify a mechanism to negotiate compression capabilities, based on scenario requirements and media types.

**2. Discussion**

Regarding the investigation for compression and impact on latency, the evaluation TR 26.847 has:

* documented model data compression using NNC.
* Documented intermediate data size with different compression profile and performances of different codecs or compression technologies (e.g. NNC, Quantization)

Further, TR 26.927, FCM is documented as a candidate technology along with its performances for vision-based scenarios and corresponding AIML models.

A new contribution (XXX) demonstrates that a complete AI split inferencing system can be achieved with existing tools and codecs to achieve acceptable latencies and results for a given frame rate.

**3. Proposal**

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

\* \* \* Begin 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, and resulting in an output which may be a processed version of the media (e.g. picture enhancement, audio translation) , or a specific description of the input media itself, such as 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, 3 approaches have been documented:

* UE devices with sufficient AI capabilities may support on device AI inferencing.
* Support for AI inferencing in the network for use cases where on device AI inferencing may be difficult or infeasible.
* Split AI processing between UEs and servers to balance the resources, power consumption, efficiency and privacy considerations.

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 splitAI/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 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 AI/ML model distribution, split AI/ML operation and distributed/federated learning. Different AI user plane data components have been identified and documented (i.e. 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 describes 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:

Possible normative work

- For collaboration scenarios 1 (Over The top) and 2 (Hosting), specify and extend the relevant procedures to support the delivery of AI data components in the architectures of TS 26.501 and TS 26.506 for 5GMS and RTC respectively.

- For collaboration scenario 3 (MNO-operated), specify and extend 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, in the architectures of TS 26.501/TS 26.506 and TS 23.228 for 5GMS/RTC and IMS DC respectively

For the above scenarios:

- Specify the use and potential extension of existing APIs at referencing points to support the delivery of AI data components.

- Specify 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.

- Define the 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

- Whenever possible, specify one or more 3GPP interoperable formats (e.g. ONNX) for the AI data components associated with the relevant AI media services

-

-

-

- Specify the mechanism to address the necessary and different interoperable compression profiles 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.

Possible further study:

[

- 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, notably relevant to collaboration scenario 3, including distributed/federated learning

]

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