**3GPP TSG-SA WG4 Meeting #130S4-241954Interdigital**

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

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

**Title: [FS\_AI4Media] pCR on data components**

**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 \* \* \* \*

## 6.1 General

Based on the identified use cases in clause 4.2, this clause documents the different data components involved in AI/ML-based media services. The delivery of certain data components in either the downlink or uplink direction between the UE and the network is dependent on both the use case requirements and service configuration.

AI data as handled by the AI Data Access/Delivery function consists of AI model data (the data representing trained or untrained models, as well as their associated parameters/weights), and in the case of split AI/ML inferencing, AI intermediate data (the data output from the endpoint performing a first split inference, typically sent as the data input into the endpoint performing a second split inference).

The data formats for AI data components is dependent on the framework used; existing frameworks such as TensorFlow and PyTorch are popular frameworks in both the tech industry and academia. ONNX provides an interoperable AI model format with a uniform model representation to facilitate the exchange of machine learning models between different other AI frameworks (Pytorch, TensorFlow). ONNX is the de-facto standard for this purpose.

A training phase and an inference phase can be decoupled by adopting ONNX. ONNX can be used to convert a trained model (e.g. Pytorch or TensorFlow) for inference in an ONNX runtime adapted for deployment target.

Clause 6.6 also introduces certain types of metadata which may be relevant to AI/ML-based media services, namely that related to AI model information, split AI/ML operations and service requirement/endpoint capability.

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