# 

**3GPP TSG-SA WG4 Meeting 132  *S4-250922***

**Fukuoka, JP, 19 – 23 May 2025**

**Agenda item:** 9.6

**Source:** Huawei, HiSilicon

**Title:** [FS\_AI4Media] Suggested update to the conclusions

**Version:** 1.1.1

**Document for Discussion and Agreement**

# Discussion

The study presented in 26.927 is rather extensive and covers many different aspects, starting from SA1 cases

1. Model split
2. Model distribution
3. Federated learning

SA4 media related cases are introduced (object recognition, video enhancement, crowd source, NLP on speech processing) showing clear distinction that in these use cases there is:

1. Processing/inferencing at UE (with or without prior model distribution)
2. Processing/inferencing in the network
3. Split operation distributing model operations.

In addition different architectures are covered

* Generalized architecture
* 5GMS/RTC
* IMS based

Last but not least, the data models and formats available to implement compression/transmission etc are covered in clause 6.

The problem is that for many outside of SA4 they will look only at the conclusions, currently the text in the conclusion does not very specifically distinguish all these aspect and is not self-contained. Therefore some text updates are proposed to improve the structure and make the conclusion more self-contained and possibly pointing back into the report.

This PCR is an initial attempt to improve this and structure a bit more providing a bit more structured summary of the content of the report. Non of the conclusion recommendations are altered in the PCR.

# Proposal

Review the text proposal and agree on updating the conclusion text to 26.927

|  |
| --- |
| \*\* CHANGES \*\* |

# 7 Conclusion

AI/ML in media services involve the use of AI/ML models to perform media processing. The AI/ML models used to process media, typically involving video, audio as input. The output can be an enhanced version of the media, such as improved picture quality or translated audio, a detailed description of the media itself, like object recognition labels, or entirely new media, such as converting text or sign language into speech or video. In order to support such AI/ML based media processing, three common scenarios have been studied in the context of different use cases and architectures:

- UE device AI inferencing

- AI inferencing in the network

- Split AI inferencing between the UE and the network.

The following media related use cases were considered:

* Object detection, see clause 4.2.2
* Video quality enhancement in streaming , see clause 4.2.3
* Crowd source media capture, see clause 4.2.4
* NLP on speech, see clause 4.2.5

In this study, the broad findings for AI/ML model transfer in TR 22.874 [2] have been further analysed with specific focus on media-based AI/ML use cases and scenarios. In particular:

* Model split operation in the 5G system and IMS for media related use cases.
* Model delivery in the 5G system and IMS for media related use cases.
* Federated learning in the 5G system and IMS for media related use cases.

This document described how AI/ML models and data may be distributed over the 5G system and/or IMS and documents the split AI/ML operations between different AI/ML endpoints (noticeably the UE and the network). This includes some study of compression of AI model data and intermediate data due to the broad range of applications and AI models feasibility evaluations for a given set of scenarios are documented in TR 26.847 [20] as part of this study.

Functional architectures are presented for:

* Basic AI/ML model distribution,
* Split AI/ML operation
* 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 including respective procedures. This mapping outlines how AI media use cases integrate with various architectures and procedures, detailing the provisioning, capability discovery/negotiation, and session support for delivering AI data components. It also explains the use of AI media functions at different endpoints based on the negotiated service configuration. Additionally, three collaboration scenarios are studied, each offering varying levels of MNO network support for AI/ML functions Clause 6 details data formats for AI components that can possibly be used to implement the different architecture.

and its related, , these

Based on the details in the report, the following next steps are identified:

Normative work in release-20:

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 in IMS services, as identified in clause 5.4.

- Extend TS 26.114 and TS 26.264 specifications to support AI/ML media processing in IMS services, as identified in clause 5.5.

- Specify support for AI/ML data signalling and negotiation, including support for split.

- Select interoperable formats for AI/ML model data and intermediate data.

- Define the support of the configuration, delivery, compression, and processing of AI/ML data in a new specification, as needed.

- Collaborate with SA2 on related matter where necessary.

New Study in Release 20 or beyond:

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, metrics and/or QoS procedures to support the delivery of the 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.

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:

- The support of AI/ML model distribution and operation, based on details in clause 5.3.6.

- The support of split AI/ML inferencing between the UE and the network, based on details in clause 5.3.5.

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

- Collaborate with SA2 on related matter where necessary for Release 20 or beyond.