



PoC 21: Validation of
LLM for Network OAM
Application on Generic
Computing Platform

Rapporteur: China Telecommunications, Intel, China Unicom, Huawei (UK) Technologies Ltd., CAICT



#### PoC Goals and PoC member task

**Host/Team Leader:** 



- User Stories / Use Cases definitionPoC development, documentation, and demos

Team members:



-Help with the architecture design, implementation of algorithm, testbed setup





-Participation in project discussions



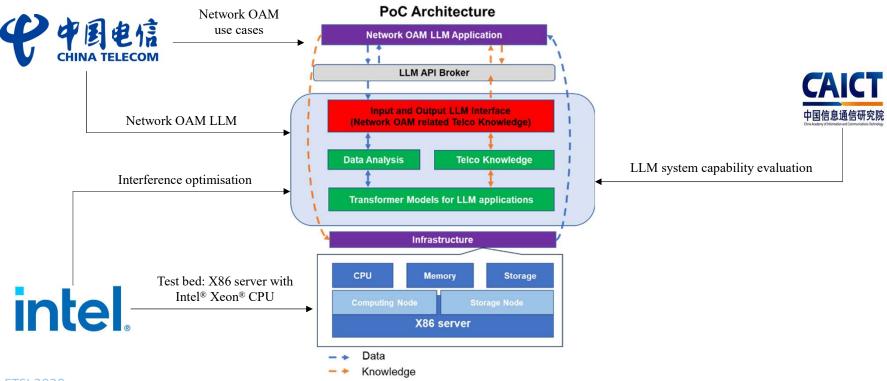
- LLM system capability evaluation

- ✓ PoC Project Goal #1: AIGC application on X86 Platform. Demonstrate how to support AIGC application on X86 platform, support AIGC inference for LLM services, and realize the functionality and capability to provide identical AIGC request.
- ✓ PoC Project Goal #2: Distributed AIGC computing optimization. Demonstrate organizing multiple computing node to provide a distributed solution for AIGC application.



#### **PoC Architecture**

The diagram represented below shows the framework of the PoC mapping to the ENI reference architecture.





**Background:** This PoC intends to demonstrate the feasibility and capability of network OAM LLM application running on generic computing platform aka X86 based platform instead of a GPU platform, with special attention to the lower cost and power consumption aspects, in the context defined by ENI.

In particular, this PoC solves the adaptation of generic computing platform e.g. X86 to replace part or all of a GPU platform for a reduced power and cost consumption. As demand for GPU increase dramatically due to the rise of AIGC applications, service providers having difficulties to provide sufficient computing power for AIGC related applications. The X86 platform on the other hand, cumulated large amoung of computing power by stable service provider investment. The overall CPU workload for service provider often very low, less than 50%. To solve the unmatched problem of computing power and demand, it is a possible solution to employ AIGC applications on X86 servers. This PoC will demonstrate the computing ability by X86 platform, which realizes the balance between the CPU workload and lowering the power and budget cost.



### Testbed Setup: X86 server with Intel® Xeon® CPU for LLM inference

СРИ	5 <sup>th</sup> Gen Intel <sup>®</sup> Xeon <sup>®</sup> (Emerald Rapids)
Total Cores	52
Total Threads	104
Max Turbo Frequency	3.1Ghz
Processor Base Frequency	2.6 GHz
Cache	260MB

### Test results with China Telecom's Network OAM LLM (13B)

I/O=1024/128				
Precision	First Token	Next Token	Throughput	
BF16/FP16	540.75 ms	48.19 ms	20.75 tokens/s	

question:请列举出中国电信相关的规章制度,包括但不限于但不限于:操作规程、应急预案、培训、设备维护保养、检查等。 <s> <human : 请列举出中国电信相关的规章制度,包括但不限于但不限于:操作规程、应急预案、培训、设备维护保养、检查等。 <bot> :以下是中国电信相关的规章制度:

- 1. 操作规程: 如设备安装/退网、线路施工/维护、网络调整、故障处理等。
- 2. 应急预案: 如自然灾害、网络事件、通信阻断、设备故障等。
- 3. 培训:包括职业安全、法律法规、产品技能、管理技能等方面的培训。
- 4. 设备维护保养:设备的维护规范、保养流程、故障处理等。
- 5. 检查:包括现场安全检查、网络运行质量检查、现场环境检查等。</s>



### **PoC Progress**

China Telecom and Intel to release a white paper in workshop jointly organized by CCSA TC610 and ETSI ENI 13<sup>th</sup> September 2024 on the x86 based inference solution for network LLM

The PoC related project also entering the final round in "Huacai" computational power national wide competition, results to release on 13<sup>th</sup> September 2024

PoC solution also passed a CAICT certification test for the computational power capability.









大语言模型(LargeLanguage Model,LLM,以下周称"大概型")作为人工智能(Anticia httelagence,A)划划的外极性技术,正凭展现致分数多形成为现长化的。为各许空 编成强力支持。作为中国领先的电话运营商之一,中国电话也统托其丰富的业务场册 需求和雇助则需数据与知识,自主或并指出信息通信物的宣言。同格大模型—— 中国电话网络大模型,用以源升其云阔环境下的网络自智水平。

目前,中国电信周绪大概型已在中国电信规则各省公司实现规模落地,并深度嵌入割中国电信集团及省 (A (移动器)、中国电信大规型平台 (Web 物) 等联网系统,实成全国部署。建立大小模型协同、促进合作自1等包提升,概题省公司—进入规则成成编模。 流少入为错误,解决现其实际问题,为全用运修/运费人员提供相应能力,使运动收率需要的心心。 原本规划电阻。







© ETSI 2020

W251-71257



### **PoC Milestones and Current Progress**



PoC Milestone	Stages/Milestone description	Target Date	Additional Info
P.S	PoC project submission	03/2024	Proposal approved for PoC reviewing during #ENI 29
P.D1	PoC Demo 1	06/2024	Demo in ENI #30
P.D2	PoC Demo 2	09/2024	Demo at ENI#31
P.C1	PoC Expected Contribution 1	10/2024	contributions to ENI requirements.
P.C2	PoC Expected Contribution 2	10/2024	contributions to ENI use case.
P.R	PoC Report	12/2024	PoC-Project-End Feedback
P.E	PoC Project End	01/2025	Presented to ISG ENI for information
P.L	r oo r roject End	01/2023	1 reserved to 100 ENT for information