



6G and Generative AI: Co-Creating the Future Intelligence

2025 6G White Paper





6G and Generative AI: Co-Creating the Future Intelligence

2025 6G White Paper

6G and Generative AI: Co-Creating the Future Intelligence

발행및편집인 장경희(6G포럼 집행위원장)

출판등록 2025년 5월 15일

주소 서울 중구 무교로 32, 효령빌딩 307호

전화 02-6248-3503

홈페이지 www.6Gforum.or.kr

디자인 그리니(Griny)

ISBN 979-11-94751-05-2 95560 (PDF)
〈비매품/무료〉

본 저작물은 2025년 6G포럼 서비스위원회에서 작성하여 6G 포럼
홈페이지(www.6Gforum.or.kr)에서 무료로 다운받으실 수 있으며,
저작권법에 따른 공공누리(공공저작물 자유이용 허락 표시제도)
제2유형(출처표시, 상업적 이용금지의 보호)를 받고 있습니다.

Notification

Copyright © 2025 by 6G Forum.
All rights reserved



목 차



약 어	1
1. 서론	5
2. GenAI 서비스 및 기술 동향	11
2.1. GenAI 개요	13
2.2. GenAI 서비스 사례	14
2.3. GenAI 산업 생태계: 기술, 인프라, 응용 서비스의 융합	16
2.4. GenAI 관련 산업체별 동향	19
2.4.1. Google	19
2.4.2. Apple	21
2.4.3. OpenAI	22
2.5. Agentic AI와 Physical AI	24



3. AI-RAN 서비스 및 기술 동향 27

3.1. AI-RAN 필요성	29
3.1.1. 비즈니스 관점	29
3.1.2. 기술 관점	30
3.2. AI-RAN 개요	31
3.3. AI-on-RAN	35
3.3.1. AI/ML 모델 분할 기반 스펙트럼 센싱	36
3.3.2. AI 기반 통신 및 센싱 결합	38
3.4. AI-for-RAN	40
3.4.1. AI 기반 상향링크 채널 보간법	41
3.4.2. AI 기반 PUSCH 채널 추정 기법	42
3.4.3 모빌리티 인식 AI 기반 간섭 완화 및 에너지 절약을 위한 5G 빔포밍 기법	44

4. 6G 서비스 및 기술 동향 47

4.1. 위성통신 서비스	52
4.2. 로봇 서비스	54
4.3. 디지털 트윈 서비스	56
4.4. 메타버스 서비스	58
4.5. 스마트 헬스케어 서비스	60
4.6. 자율주행 및 모빌리티 서비스	62
4.7. 스마트시티 및 사물인터넷 서비스	64



4.8. 산업자동화 서비스	66
4.9. 홀로그램 서비스	68
5. 6G와 GenAI 융합 미래 모바일 서비스.....	71
5.1. GenAI-on-6G 서비스 시나리오	73
5.1.1. 통신사 대규모 언어모델 및 AI 에이전트 서비스	73
5.1.2. 통신사 무선 접속망과 GenAI를 결합한 지역 GPT 서비스	79
5.1.3. GenAI 기반 ISAC 융합 긴급 대응 서비스	82
5.1.4. 엣지-클라우드 기반 GenAI·로봇·드론 융합 농업 서비스	83
5.1.5. 피싱 범죄 예방을 위한 On Device Computing과 GenAI 활용 서비스	85
5.1.6. GenAI를 활용한 지능형 도심 항공 모빌리티 서비스	86
5.2. GenAI-for-6G 서비스 시나리오.....	90
5.2.1. GenAI 기반 디지털 트윈을 이용한 네트워크 최적화 서비스	90
5.2.2. 텍스트-이미지 생성을 위한 대규모 언어모델 기반 시맨틱 통신 서비스	92
5.2.3. 스마트시티에서 GenAI와 디지털 트윈을 활용한 네트워크 에너지 소모 최적화 서비스	94
6. 결론	99
참고문헌	103
백서 편집위원회	110



그림목차



[그림 2-1] Gemini 1.5와 타 모델 간 비교	20
[그림 2-2] Image Playground 예시.....	22
[그림 2-3] 다양한 업무에서의 모델 간 성능 비교	24
[그림 3-1] 통신사의 총수익 변화.....	29
[그림 3-2] AI-RAN을 활용한 Telco의 비즈니스 변화	30
[그림 3-3] AI 와 RAN의 통합.....	31
[그림 3-4] AI-RAN Alliance의 Founding member.....	32
[그림 3-5] AI-RAN Alliance의 General member	33
[그림 3-6] AI와 RAN의 결합방식 분류.....	33
[그림 3-7] AI-on-RAN의 구조	35
[그림 3-8] 동적 AI/ML 모델 분할기반 스펙트럼 센싱 구조.....	37
[그림 3-9] 동적 AI/ML 모델 분할기반 스펙트럼 센싱의 단말 GUI 화면.....	38
[그림 3-10] AI 기반 통신 및 센싱 결합 시나리오	39
[그림 3-11] AI 기반 통신 및 센싱 결합 결과.....	39
[그림 3-12] AI-for-RAN 구조	40
[그림 3-13] AI 기반 UL Channel Interpolation.....	42
[그림 3-14] AI 기반 PUSCH 채널 추정 테스트베드 구조	43
[그림 3-15] PUSCH Throughput 결과	44
[그림 3-16] RSG 기반 AI-RAN training 및 UE mobility patterns	45
[그림 4-1] 6G 서비스 및 기술 동향 구성도	51
[그림 4-2] 위성통신 서비스 구성도	52
[그림 4-3] 로봇 서비스 활용 예	54
[그림 4-4] 디지털 트윈 서비스 개념도	56
[그림 4-5] 메타버스 서비스 개념도	58
[그림 4-6] 스마트 헬스케어 서비스 구성도	60



[그림 4-7] 자율주행 및 모빌리티 서비스 구성도	62
[그림 4-8] 사물인터넷 기반 스마트시티 구성도	64
[그림 4-9] 산업자동화 서비스 구성도.....	66
[그림 4-10] 훌로그램 서비스 활용 예.....	68
[그림 5-1] Telco LLM의 서비스 예시	73
[그림 5-2] LG Uplus의 통신 특화 언어모델 익시전의 특징	75
[그림 5-3] LG Uplus의 통신 특화 언어모델 익시전의 학습 단계	75
[그림 5-4] LG 유플러스의 AI 통화 에이전트 ixi-O의 기능	76
[그림 5-5] KT 밀:음 서비스 구성	77
[그림 5-6] SKT 에이닷엑스(A.X) 구축과정	78
[그림 5-7] 통신사 6G 네트워크와 GenAI 및 무선 접속망 기반 한국형 GPT 서비스	79
[그림 5-8] GenAI와 ISAC 기반 상호작용형 긴급 대응 서비스	82
[그림 5-9] 6G와 GenAI·로봇·드론을 활용한 미래형 농업 서비스	83
[그림 5-10] GenAI를 활용한 보이스 피싱 감지 서비스.....	85
[그림 5-11] 지능적 도심 항공 교통(Intelligent Urban Air Mobility) 서비스.....	86
[그림 5-12] GenAI 기반 디지털 트윈을 이용한 6G 네트워크 최적화 구조 및 과정.....	90
[그림 5-13] LLM 기반 시맨틱 통신 시스템 모델.....	93
[그림 5-14] 스마트시티를 위한 GenAI 기반 디지털 트윈.....	94



표목차

[표 2-1] 기술 개발 분야 주요 플레이어 및 역할.....	17
[표 2-2] 인프라 제공 분야 주요 플레이어 및 역할	18
[표 2-3] 응용 서비스 분야 주요 플레이어 및 역할	18
[표 3-1] 기존 RAN의 기술적 한계.....	30
[표 4-1] 위성통신 서비스의 활용 사례.....	53
[표 4-2] 로봇 서비스의 활용 사례.....	55
[표 4-3] 디지털 트윈 서비스의 활용 사례.....	57
[표 4-4] 메타버스 서비스의 활용 사례.....	59
[표 4-5] 스마트 헬스케어 서비스의 활용 사례.....	61
[표 4-6] 자율주행 및 모빌리티 서비스의 활용 사례	63
[표 4-7] 스마트시티 및 사물인터넷 서비스의 활용 사례	65
[표 4-8] 산업자동화 서비스의 활용 사례.....	67
[표 4-9] 훌로그램 서비스의 활용 사례.....	69

Abbreviations





Abbreviations



AI	Artificial Intelligence	LOS	Line-Of-Sight
AR	Augmented Reality	MAC	Medium Access Control
COTS	Commercial Off-The-Shelf	MEC	Mobile Edge Computing
CQI	Channel Quality Indicator	ML	Machine Learning
CSI	Channel State Information	MMSE	Minimum Mean Square Error
CNN	Convolution Neural Network	MR	Mixed Reality
DL	DownLink	mmWave	millimeter Wave
D-MIMO	Distributed Multiple-Input Multiple-Output	mMTC	massive Machine-Type Communications
eMBB	enhanced Mobile BroadBand	NLOS	Non-Line-Of-Sight
FWA	Fixed Wireless Access	NLP	Natural Language Processing
GAN	Generative Adversarial Networks	NN	Neural Network
IoT	Internet-of-Things	NTN	Non-Terrestrial Network
ISAC	Integrated Sensing And Communications	PBCH	Physical Broadcast CHannel
ITU	International Telecommunication Union	PUSCH	Physical Uplink Shared CHannel
JCAS	Joint Communication And Sensing	QoE	Quality of Experience
LEO	Low-Earth Orbit	RAN	Radio Access Network
LLM	Large Language Models	RIC	RAN Intelligent Controller



Abbreviations

RL	Reinforcement Learning
RLHF	Reinforcement Learning based on Human Feedback
RSG	RAN Scenario Generator
RSRP	Reference Signal Received Power
RSSINR	Reference Signal Strength Indicator to Inference-plus-Noise Ratio
rApp	RIC Application
SNR	Signal Noise Ratio
SRS	Sounding Reference Signal
UE	User Equipment
UL	UpLink
UAM	Urban Air Mobility
URLLC	Ultra-Reliable and Low-Latency Communications
V2I	Vehicle-to-Infrastructure
V2V	Vehicle-to-Vehicle
V2X	Vehicle-to-Everything
VR	Virtual Reality
vRAN	virtual RAN
XL-MIMO	eXtra-Large Multiple-Input Multiple-Output
xApp	external Application

1. 서론





1. 서론

기술의 진보는 인간의 상상력을 현실로 구현하는 방향으로 끊임없이 진화하고 있다. 그 중심에는 6G 이동통신과 생성형 인공지능(Generative AI, GenAI)이라는 두 가지 핵심 기술이 있다. 각각 독자적으로 발전해 온 이 두 기술은 이제 서로 결합하며, 미래 지능 사회(Future Intelligence)를 공동으로 창조하려는 새로운 전환점을 맞이하고 있다. 6G는 전례 없는 통신 인프라 혁신을 이끌며, 데이터 중심 사회의 기반을 마련할 전망이다. 한편, GenAI는 대규모 데이터 학습을 통해 인간의 창의력을 보완하고 지능형 서비스를 자동으로 생성하는 기술로, 다양한 산업 분야와 일상 생활의 주요 영역에서 빠르게 활용되고 있다. 이 두 기술의 융합은 단순한 연결을 넘어, 인간과 기계, 물리와 디지털, 네트워크와 지능의 경계를 허물고 새로운 형태의 지능형 사회를 구현하는 원동력이 될 것이다. 6G와 GenAI는 상호 보완적으로 작용하며 미래 사회가 요구하는 복잡하고 동적인 문제를 해결하는 데 필수적인 역할을 수행할 것으로 기대된다.

6G는 5G를 넘어 진화하는 차세대 이동통신 기술로, 미래 사회의 초연결성과 초지능화 요구를 충족하기 위해 개발되고 있다. International Telecommunication Union (ITU)은 6G를 위한 비전으로 인공지능 및 통신 융합(Integrated AI and Communication), 초고신뢰·초저지연 통신(Hyper Reliable and Low-latency Communication), 유비쿼터스 연결(Ubiquitous Connectivity), 대규모 통신(Massive Communication), 통합 감지 및 통신(Integrated Sensing and Communication, ISAC), 몰입형 통신(Immersive Communication) 등 여섯 가지 주요 사용 시나리오를 제시하였다. 이러한 6G의 비전은 단순히 기존 세대의 통신 성능을 확장하는 것을 넘어, 지능형 서비스와 물리적 세계를 실시간으로 연결하는 새로운 디지털 인프라를 지향한다. 3GPP를 포함한 주요 표준화 기구에서는 6G 초기 단계에서 인공지능 기반 무선 인터페이스, 에너지 효율성 향상, 네트워크 자율 최적화 기술 등을 중심으로 표준화 논의를 진행하고 있다. 특히, AI 기술은 6G 네트워크의 본질적인 구성 요소로 자리매김하고 있으며, AI-Native 네트워크 구현을 포함하여 다양한 형태로 6G 기술 발전을 지원하고 있다. 6G는 초저지연성과 초고속 전송 능력을 통해 다양한 미래형 서비스를 구현하는 핵심 인프라로 활용되어, 향후 여러 산업 분야에서 디지털 전환을 가속하는 기반이 될 것으로 전망된다.

GenAI는 대규모 데이터 학습을 기반으로 텍스트, 이미지, 음성, 동영상 등 다양한 형태의 콘텐츠를 자동으로 생성하는 AI 기술이다. GenAI는 단순한 정보 제공을 넘어, 스스로 패턴을 학습하고 새로운 콘텐츠를 창출함으로써 인간의 창의력을 보완하고 산업 전반에서 생산성과 효율성을 향상시키는 핵심 기술로 자리 잡고 있다. 기술적으로 GenAI는 대규모 언어 모델(Large Language Model, LLM), 생성적 적대 신경망(Generative Adversarial Network, GAN), 멀티모달 학습 모델 등 다양한 딥러닝 기반 접근법을 활용하며, 자연어 처리(Natural Language Processing, NLP), 컴퓨터 비전, 음성 합성 등 여러 AI



기술이 융합되어 구현된다. 특히 멀티모달 모델은 텍스트, 이미지, 음성 데이터를 통합적으로 처리하여 보다 복합적이고 정교한 결과물을 생성할 수 있도록 지원한다. GenAI 기술은 콘텐츠 제작, 고객 서비스, 의료, 금융, 교육, 엔터테인먼트 등 다양한 산업 분야에서 빠르게 확산되고 있으며, 글로벌 기술 기업들의 적극적인 투자와 기술 발전에 따라 향후 산업 전반에서 새로운 비즈니스 모델과 가치 창출을 이끌어낼 것으로 예상된다.

6G와 GenAI는 상호 강화적인 관계를 통해 서로의 발전을 촉진하는 핵심적인 기술로 자리매김하고 있다. GenAI의 고도화된 개인화 기능과 실시간 대응 능력을 충분히 활용하기 위해서는 초저지연, 초고속 데이터 전송, 그리고 지능형 컴퓨팅이 필수적인데, 이러한 요구를 충족할 수 있는 최적의 인프라가 바로 6G 네트워크이다. 반대로, GenAI는 6G 네트워크의 자율 운영, 트래픽 예측, 자원 최적화 등 네트워크 자체의 지능화를 가능하게 하고, 이를 통해 6G의 기술적 완성도를 더욱 높이는 역할을 수행할 수 있다. 결과적으로, 6G는 GenAI를 위한 기술적 기반을 제공함으로써 GenAI 서비스의 범위와 가능성을 확장시키며, GenAI는 6G 인프라를 기반으로 새로운 형태의 맞춤형, 몰입형 서비스를 구현하여 6G 네트워크의 활용 가치를 극대화할 수 있다. 이러한 상호 강화적 관계를 바탕으로 6G와 GenAI의 융합은 미래 디지털 사회의 구체적인 청사진을 제시하고, 다양한 산업 분야에서 혁신적인 서비스 모델을 실현하는 핵심 동력이 될 것으로 기대된다.

본 백서에서는 6G와 GenAI의 융합이 가져올 미래 사회와 산업의 변화를 심층적으로 전망하고, 그 과정에서 등장할 다양한 서비스 모델과 기술적 도전 과제를 분석하고자 한다. 이러한 분석은 최신 연구 동향과 실제 응용 사례를 기반으로 하여, 기술적 접근법과 전략적 방향성을 체계적으로 모색하는 데 중점을 두고 있다. 특히 위성통신, 로봇 서비스, 디지털 트윈, 메타버스, 스마트 헬스케어, 자율주행, 스마트시티, 산업 자동화, 훌로그램 서비스 등 주요 산업 분야에서 실현 가능한 서비스 모델을 구체적으로 제시하고, 차세대 디지털 사회가 요구하는 기술 발전 방향과 글로벌 경쟁력을 강화하기 위한 실질적인 전략을 탐구한다. 더불어 본 백서는 6G와 GenAI 융합 기술의 성공적인 구현 방안을 탐구하기 위해, 다양한 연구·개발 사례를 통해 기술의 실현 가능성을 검토하고자 한다. 최신 기술 동향을 반영하여 사회적·경제적 측면에서 기술 발전이 새로운 가치 창출로 이어질 수 있는 가능성을 모색하며, 정책적 제도 마련과 산업 전반의 협력 체계 구축의 필요성을 재고한다. 이를 통해 6G와 GenAI가 공동으로 구축하게 될 미래 지능형 사회의 구체적인 청사진을 제공하고자 한다.

1장에 이은 본 백서의 구성은 다음과 같이 6G와 GenAI 융합 기술의 동향과 응용 가능성을 체계적으로 다룬다. 2장에서는 GenAI의 개념, 기술 구성 요소, 산업 생태계 및 주요 서비스 사례를 중심으로 기술적 특성과 산업적 활용 가능성을 분석한다. 3장에서는 AI 기술을 무선 접속망에 적용하는 AI-RAN의 개념과 역할을 설명하고, AI-on-RAN 및 AI-for-RAN 구조에 따른 주요 기술 동향과 응용 방안을 소개한다. 4장에서는 6G의 주요 비전과 핵심 기술을 정리하고, 다양한 응용 분야별 6G 기반 서비스 모델을 제시한다.



5장에서는 GenAI-on-6G 및 GenAI-for-6G 관점에서의 기존 연구·개발 사례를 통해 기술적·산업적 가능성을 조망하며 6G와 GenAI의 융합을 기반으로 한 미래 서비스 시나리오를 탐구한다. 마지막으로 6장에서는 6G와 GenAI 융합 기술의 성공적인 도입과 지속 가능한 발전을 위한 기반 마련의 중요성을 논하며 본 백서의 결론을 맺는다.



2. GenAI 서비스 및 기술 동향





2. GenAI 서비스 및 기술 동향

2.1. GenAI 개요

GenAI는 텍스트, 이미지, 오디오, 비디오 등 다양한 형태의 콘텐츠와 아이디어를 스스로 생성해낼 수 있는 인공지능을 의미한다. 이 기술은 2010년대 주류를 이루었던 분류(Classification)를 기반으로 하는 분석 및 예측 AI와는 뚜렷한 차이가 있으며, 인공지능이 인간이 만든 데이터의 패턴을 학습하여 유사하면서도 독창적인 콘텐츠를 효율적으로 창작해낼 수 있다는 개념에 기반하고 있다. GenAI를 활용하면 지금까지는 기계가 흉내내기 어려웠던, 인간의 고유영역이라 생각되어온 창의적인 작업 영역에서 인공지능이 인간을 대신하여 새로운 결과물을 만들어낼 수 있다는 점에서 다양한 산업 분야에서 혁신적인 기회가 창출될 것으로 기대된다.

GenAI의 급격한 발전은 Transformer 아키텍처를 기반으로 한 GPT(Generative Pre-trained Transformer) 모델의 등장을 기점으로 본격화되었다. 특히 2020년 5월 발표된 GPT-3는 이전 모델 대비 압도적인 1,750억 개의 파라미터를 학습하며 모델 규모의 확장에 따라 성능이 현격하게 향상될 수 있음을 입증했고, 이후 빅테크 기업들을 중심으로 대규모 언어 모델(Large Language Model, LLM) 기반의 GenAI 개발 경쟁과 LLM의 학습에 필수적인 GPU 확보 경쟁이 촉발되었다.

GenAI가 대중적인 관심과 인기를 받게 된 시점은 일반인이 사용하기 편리하도록 대화형 인터페이스에 최적화된 ChatGPT가 발표된 2022년 11월이며, 이 때부터는 ChatGPT와의 대화가 튜링 테스트를 통과하는 등 인간과 구별이 어렵고, 심지어는 자의식이 있는 것으로 느껴진다는 수준에까지 이르렀다. 2025년 봄에는 GPT-4o에서 인물 사진을 지브리 스타일로 변환하여 소셜미디어에 업로드하는 것이 대유행 할 정도로 콘텐츠 변환이 간단하고 빠를 뿐만 아니라 품질도 훌륭한 단계에 이르렀다.

이러한 큰 변화를 가져온 것은 Attention 메커니즘을 기반으로 한 Transformer 아키텍처로서, 자연어 처리 기술을 활용하여 인간이 작성한 다양한 언어 기록을 학습하고 텍스트를 생성하거나 정보를 요약하는 역할을 수행한다. Stable Diffusion과 GAN 등의 이미지 생성 기술도 고품질의 시각 콘텐츠를 제작하는 데 중요한 기반 기술이다. 또한 텍스트, 이미지, 음성 등의 데이터를 통합적으로 학습하고 처리하는 멀티모달 기능도 사용자에게 더욱 풍부하고 다채로운 경험을 제공할 수 있게 된 주요 요소이다.

이러한 GenAI의 특징으로부터 서비스에 중요한 측면 두 가지를 다음과 같이 생각해볼 수 있다. 첫째, GenAI는 글짓기와 작곡 같은 인간의 창조적 활동을 신속하게 모방할 수 있다는 점이다. 둘째, GenAI는 인간과 소통하기 위해 인간에게 친숙한 대화형 인터페이스를 그대로 사용할 수 있다는 점이다. 이를 특징에



2. GenAI 서비스 및 기술 동향

더하여, 개개인의 데이터를 GenAI가 활용할 수 있다면 개인별 맞춤형 서비스도 용이하게 생성해낼 수 있다. 즉, 이전의 개인별 맞춤형 서비스가 개개인의 성향을 분류하고 범주화하여 제공된 것이었다면, GenAI의 개인별 맞춤형 서비스는 개인의 정보와 특성에서 직접 생성될 수 있다. 고성능의 컴퓨팅 인프라 환경에서는 실시간에 가까운 속도로 결과물을 생성해낼 수도 있어, GenAI에 기반한 실시간 대화형 사용자 인터페이스(User Interface)도 충분히 예상 가능하다. GenAI가 가진 이러한 특성들로 광고, 영화, 게임 등의 콘텐츠 제작 과정을 혁신적으로 변화시킬 수 있으며, 창의성이 필요한 교육, 의료, 디자인 및 고객 서비스 등에서도 큰 변화를 예상할 수 있다.

그러나 GenAI 기술이 보다 널리 사용되기 위해서는 몇 가지 해결해야 하는 과제가 있는 것도 사실이다. 부정확한 답변을 제시하는 환각 현상, 무분별한 데이터 학습으로 인한 저작권 침해, 학습에 사용되는 데이터의 품질로 인한 콘텐츠의 편향성, 딥페이크 등 위조 콘텐츠를 이용한 사기 행위 등이 대표적이다. GenAI 모델의 결과를 신뢰할 수 있도록 인간의 개입을 전제하는 Human-in-the-Loop과 같은 안전 장치에 대한 논의가 활발히 이루어지고 있다. 또한, 대규모 AI 모델을 학습하고 실행하는 데 드는 막대한 컴퓨팅 자원과 그에 따른 에너지 소비 역시 해결해야 할 주요 과제 중 하나이다.

2.2. GenAI 서비스 사례

GenAI는 대규모 데이터를 학습한 딥러닝 모델을 기반으로 텍스트, 이미지, 음성, 동영상 등의 콘텐츠를 생성할 수 있는 혁신적인 기술이다. 이 기술은 창의성, 효율성, 맞춤형 솔루션을 제공하며, 다양한 산업에서 실질적인 가치를 창출하고 있다. GenAI 서비스는 콘텐츠 제작, 고객 서비스, 의료, 금융, 교육, 엔터테인먼트 등 다양한 분야에 적용되며, 새로운 비즈니스 모델과 사용자 경험을 가능하게 한다. 본 절에서는 GenAI의 구체적인 서비스 사례와 그 효과를 중심으로 살펴보자 한다.

■ 콘텐츠 제작 분야

GenAI는 광고, 마케팅, 영상 제작, 그래픽 디자인 등 콘텐츠 중심 산업에서 폭넓게 활용될 수 있다. 특히 마케팅 분야에서는 기업이 마케팅 콘텐츠를 자동으로 작성하고, 고객 맞춤형 메시지를 생성할 수 있도록 지원한다. 예를 들어 Jasper AI와 같은 플랫폼은 블로그 게시물, 이메일 캠페인, 소셜 미디어 광고 등을 자동으로 생성함으로써 마케팅 팀의 생산성을 극대화한다. 이러한 자동화 기능을 통해 기업은 고객 데이터를 기반으로 한 맞춤형 콘텐츠를 제공하여 전환율을 높이고 사용자 참여를 증대시킬 수 있다.

영상 및 이미지 생성 분야에서 Adobe는 GenAI 기술을 Creative Cloud에 통합하여, 디자이너가 빠르고



효율적으로 그래픽과 이미지를 생성할 수 있도록 지원한다. 또한, Runway ML과 같은 도구는 복잡한 영상 편집 작업을 자동화하고 고품질의 비디오 콘텐츠를 생성하는 데 활용된다. 이러한 기술은 광고, 영화, 게임 등 다양한 분야에서 널리 사용되고 있다.

게임 개발 분야에서는 GenAI가 게임 세계를 자동으로 생성하거나 스토리라인을 작성하는 데 활용된다. NVIDIA Omniverse는 AI 기반으로 실시간 몰입형 가상 세계를 생성하며, ChatGPT는 NPC(비플레이어 캐릭터)의 대화를 생성하여 게임 플레이 경험을 개선한다.

■ 고객 서비스 및 상담

GenAI는 고객 서비스 분야에서 자연스러운 대화와 맞춤형 지원을 가능하게 하며, 사용자 경험을 혁신하고 있다. 기존의 규칙 기반 챗봇과 달리, GenAI는 자연어를 이해하고 생성하는 능력이 비약적으로 향상되어 사전에 정의하지 않은 질문에도 유연하게 대응할 수 있고, 대화의 맥락을 유지하며 지속적인 상호작용이 가능하다.

대표적인 활용 예로는 챗봇과 가상 비서를 들 수 있다. OpenAI의 GPT-4를 기반으로 한 고객 서비스 챗봇은 고객의 질문에 신속하고 정확한 답변을 제공하며, 복잡한 문제 해결을 지원한다. 이러한 서비스 챗봇은 은행, 전자상거래, 보험회사 등 다양한 산업에서도 활용되며, 연중무휴(24/7) 고객 지원 서비스를 제공하고 있다.

또한 음성 지원 서비스에서도 GenAI는 핵심 기술로 활용된다. Google Assistant, Amazon Alexa, Apple Siri와 같은 음성 비서는 GenAI를 활용하여 사용자 명령을 이해하고 맞춤형 음성 응답을 제공한다. 이와 같은 서비스는 가정, 자동차, 스마트 기기 등 다양한 환경에서 사용자와의 상호작용을 향상시키며, 일상 속에서 보다 자연스러운 경험을 제공한다.

■ 의료 및 헬스케어

GenAI는 방대한 양의 의료 데이터를 신속하고 정확하게 처리할 수 있는 능력을 바탕으로, 진료의 효율성과 정확성을 높이고 개인화된 의료 솔루션을 제공하는 것에 있어 크게 기여할 수 있다.

먼저 의료 기록 요약 부분에서 GenAI는 의료진이 환자의 기록을 빠르게 정리하고, 핵심 정보를 추출하는데 도움을 준다. 예를 들어, DeepMind는 환자 데이터를 분석하여 예측 모델을 생성하고, 치료 결정을 지원하는 솔루션을 개발하고 있다.

또한, 진단 및 치료 계획 수립에도 GenAI 기반 도구가 활용된다. AI는 X-ray, MRI 등 의료 이미지를 분석하여 질병을 진단하고 치료 계획을 수립한다. 예를 들어, PathAI는 병리 데이터를 분석하여 암과 같은 질병의 조기 발견을 가능하게 한다.



2. GenAI 서비스 및 기술 동향

정신 건강 분야에서도 GenAI는 주목받고 있다. Woebot과 같은 GenAI 기반 앱은 사용자의 정신 건강 상태를 모니터링하고, 상담 및 행동 치료를 위한 맞춤형 지원을 제공함으로써 정신적 안정과 회복을 돋운다.

■ 교육 및 학습

GenAI는 교육 현장에서 학습 자료를 자동으로 생성하고, 학습자의 수준과 필요에 맞춘 학습 경험을 제공함으로써 더욱 효과적인 교육을 가능하게 한다.

대표적인 사례로 Khan Academy는 GPT-4를 활용하여 학생들의 학습 수준과 요구에 맞춘 개인화된 교육 콘텐츠를 제공하고 있으며, 이를 통해 학생들의 학습 참여도를 높이고 학습 성과를 향상시킨다.

또한 언어 학습 분야에서도 GenAI는 효과적으로 활용되고 있다. Duolingo는 GenAI를 도입해 사용자가 언어를 배우는 과정에서 실시간 피드백과 자연스러운 대화를 제공한다. AI는 사용자의 실력 수준을 분석하고, 그에 따라 학습 계획을 자동으로 조정하여 보다 효율적인 학습을 유도한다.

■ 금융 서비스

GenAI는 금융 산업에서 데이터 분석, 보고서 작성, 리스크 평가 등 다양한 업무를 지원함으로써 전반적인 효율성을 크게 향상시키고 있다. 반복적이고 시간이 많이 소요되는 작업을 자동화함으로써 전문가들이 보다 전략적인 의사결정에 집중할 수 있도록 돋운다.

투자 보고서 작성 분야에서는 GenAI가 방대한 금융 데이터를 신속하게 분석하고, 이를 기반으로 투자 보고서를 자동 생성하는 데 활용된다. 대표적으로 Bloomberg GPT는 시장 동향을 분석하여 투자 결정을 지원하는 데 사용되며, 투자 전문가들이 실시간으로 변화하는 금융 환경에 빠르게 대응할 수 있도록 한다.

또한, 고객 상담에서도 GenAI의 활용이 증가하고 있다. 은행과 금융 기관들은 AI 기반 상담원을 도입하여 고객의 재무 상태를 정밀하게 분석하고, 이에 맞춘 맞춤형 재무 계획을 제안하고 있다. 예를 들어, Wealthfront는 이러한 기술을 통해 고객의 요구에 더욱 정교하게 대응하며, 전통적인 금융 서비스보다 향상된 사용자 경험을 제공하고 있다.

2.3. GenAI 산업 생태계: 기술, 인프라, 응용 서비스의 융합

GenAI는 대규모 데이터를 학습하여 텍스트, 이미지, 음성, 동영상 등 새로운 콘텐츠를 생성하는 인공지능 기술로, 디지털 콘텐츠 생성과 다양한 산업 혁신을 주도하고 있다. 이 기술은 자연어 처리(NLP), 컴퓨터 비전, 음성 합성, 멀티모달 통합 등 첨단 기술과 융합되어 빠르게 진화하고 있다. GenAI 산업 생태계는



기술 개발, 인프라 제공, 응용 서비스 등 다양한 가치 사슬로 구성되며, 글로벌 주요 기업과 기관들이 이 시장을 주도하고 있다. 본 절에서는 GenAI 산업 생태계의 구성 요소와 주요 플레이어들의 역할을 중심으로 살펴보자 한다.

GenAI 산업 생태계는 크게 기술 개발, 인프라 제공, 응용 서비스로 구분되며, 각 단계에서 다양한 플레이어들이 협력하여 시장을 형성하고 있다.

■ 기술 개발

GenAI의 핵심 기술은 LLM, GAN, Transformer 등으로 구성된다. 이러한 기술은 학습 알고리즘과 데이터셋의 품질에 따라 성능이 결정되며, 이를 개발하는 연구기관과 기업이 생태계의 중심 역할을 한다.

- 대표 기술: GPT 시리즈, BERT, DALL-E, WaveNet 등
- 핵심 요소: 데이터 품질, 모델 정확도, 학습 효율성, 윤리적 안전성

[표 2-1] 기술 개발 분야 주요 플레이어 및 역할

OpenAI	GenAI 분야의 선도 기업으로, GPT-4와 같은 대규모 언어 모델을 개발하였다. 자연어 처리와 텍스트 생성 기술을 통해 다양한 응용 서비스를 제공한다.
Google DeepMind	구글 산하 연구 기관으로, GenAI 기술과 강화 학습 알고리즘 개발에서 주도적인 역할을 하고 있다.
Anthropic	AI 윤리와 안전성을 강조하며, 인간 친화적인 GenAI 모델 개발에 주력하고 있다.
Hugging Face	오픈소스 기반의 NLP 및 GenAI 플랫폼을 제공하며, 다양한 기업과 연구기관이 활용하고 있다.

■ 인프라 제공

GenAI 기술은 대규모 데이터 처리와 고성능 컴퓨팅 자원이 필수적이다. 클라우드 컴퓨팅 서비스, GPU와 TPU와 같은 고성능 하드웨어, 데이터 관리 플랫폼 등이 기술 구현의 기반을 제공한다.

- 주요 플레이어: 클라우드 제공업체(AWS, Google Cloud, Microsoft Azure), 하드웨어 제조사(NVIDIA, Intel)
- 핵심 요소: 컴퓨팅 성능, 데이터 저장 및 처리 효율성, 확장성



[표 2-2] 인프라 제공 분야 주요 플레이어 및 역할

NVIDIA	고성능 GPU와 AI 전용 하드웨어를 제공하여 GenAI 기술의 학습과 실행을 지원한다.
Amazon Web Services(AWS)	클라우드 컴퓨팅을 통해 GenAI 모델 학습에 필요한 컴퓨팅 자원과 데이터 저장소를 제공한다.
Google Cloud	GenAI의 데이터 처리와 모델 학습에 필요한 강력한 클라우드 기반 솔루션을 제공한다.
Microsoft Azure	대규모 언어 모델 학습과 AI 응용 개발을 지원하는 AI 전용 클라우드 서비스를 제공한다.

■ 응용 서비스

GenAI 기술은 다양한 산업과 응용 분야에서 활용되고 있다. 디지털 콘텐츠 제작, 가상 비서, 의료 진단, 교육, 게임 개발, 금융 보고서 자동화 등 여러 산업에서 사용 사례가 등장하며, 이를 통해 최종 소비자와 기업은 새로운 가치를 창출할 수 있다.

- 주요 분야: 마케팅, 교육, 의료, 엔터테인먼트, 금융 등
- 핵심 요소: 사용자 경험, 맞춤형 서비스, 비즈니스 효율성

[표 2-3] 응용 서비스 분야 주요 플레이어 및 역할

CANVA	GenAI를 활용하여 사용자가 쉽게 이미지와 디자인 콘텐츠를 생성할 수 있는 플랫폼을 제공한다.
Adobe	GenAI 기술을 통합한 Creative Cloud를 통해 디자이너와 크리에이터가 더욱 효율적으로 작업할 수 있도록 돕는다.
Jasper AI	마케팅 콘텐츠 작성과 자동화를 지원하는 GenAI 플랫폼으로, 중소기업과 마케터들에게 인기가 높다.
Synthesia	AI 기반의 영상 콘텐츠 제작 플랫폼으로, 사용자가 손쉽게 동영상을 생성할 수 있도록 지원한다.

GenAI 산업 생태계는 기술 개발, 인프라 제공, 응용 서비스가 유기적으로 연결된 형태로 발전하고 있다. 글로벌 기술 기업들이 이 시장을 주도하고 있으며, 향후 5~10년 동안 산업 전반에서 GenAI가 필수적인



요소로 자리 잡을 것으로 예상된다. 새로운 응용 사례와 비즈니스 모델이 등장함에 따라, GenAI는 미래 디지털 경제의 핵심 동력이 될 것이다.

2.4. GenAI 관련 산업체별 동향

2024년도에 GenAI를 주도하는 글로벌 기업들에서 공개 행사 등을 통해 다양한 신기술들을 발표하였는데, 대표적인 세 기업인 Google, Apple, OpenAI에서 발표된 주요 내용들을 알아본다.

2.4.1. Google

2024년 5월 구글 I/O에서 공개된 주요 기술들은 크게 제미나이(Gemini) 모델, 생성형 미디어 모델, AI 기반 검색 혁신 등으로 구분할 수 있다. 이번 행사는 AI 기술이 일상적인 작업을 넘어 창의적 작업, 검색, 교육 등 다양한 분야에 깊이 통합되고 있음을 보여주는 중요한 이정표였다.

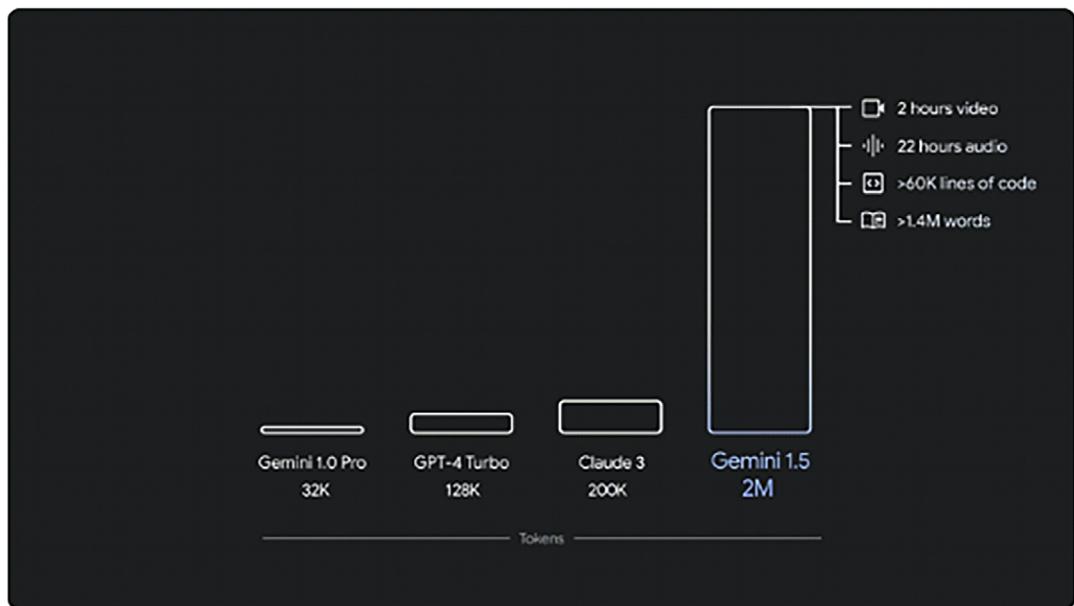
- **제미나이(Gemini) 모델:** 구글은 제미나이 1.5 시리즈를 공개하며, 특히 제미나이 1.5 플래시와 프로 모델을 강조했다. 이 모델들은 대규모 서비스에 최적화된 경량화와 성능 향상을 목표로 개발되었다. 제미나이 모델은 텍스트, 이미지, 오디오 등 멀티모달 데이터를 처리할 수 있는 능력을 갖추고 있으며, AI 스튜디오와 버텍스 AI 플랫폼에서 활용 가능하다. 또한, 제미나이 어드밴스드 모델은 대용량 문서 분석과 같은 고급 기능을 제공하며, 맞춤형 여행 일정 추천 등 개인화된 서비스를 지원한다.
- **생성형 미디어 모델:** 구글은 텍스트-이미지 변환 모델인 ‘이마젠 3(Imagen 3)’과 비디오 생성 모델 ‘비오(Veo)’를 공개했다. 이마젠 3은 복잡한 프롬프트의 세밀한 디테일까지 반영하여 사실적인 이미지를 생성할 수 있다. 비오는 1080p 해상도의 고화질 동영상을 생성하며, 유튜브와 같은 플랫폼에서 활용될 예정이다. 또한, 뮤직 AI 샌드박스(Music AI Sandbox)와 같은 창의적 도구들을 소개하여 음악 및 미디어 제작 분야에서 AI의 가능성을 확장했다.
- **AI 기반 검색 혁신:** 구글은 제미나이 모델을 통해 구글 검색의 다단계 추론 및 계획 수립 기능을 강화했다. AI 개요(AI Overview) 기능을 도입하여 복잡한 질문에 대한 답변을 한 번에 제공하며, 사용자 맞춤형 검색 결과 페이지를 생성하는 등 검색 경험을 혁신적으로 개선했다. 또한, 동영상과 이미지 기반 질문 처리 기능을 추가하여, AI로 정리된 검색 결과와 같은 새로운 검색 경험을 제공하고 있다.
- **AI와 구글 서비스 통합:** 구글은 제미나이 모델을 구글 워크스페이스와 구글 포토에 통합하여, 이메일



2. GenAI 서비스 및 기술 동향

요약, 스마트 화신, 첨부파일 분석 등 다양한 AI 기반 기능을 제공하고 있다. 또한, 안드로이드와 구글 메시지에서도 AI를 활용한 이미지 생성 및 텍스트 분석 기능이 강화되었다. 이를 통해 사용자들은 더욱 스마트하고 효율적인 작업 환경을 경험할 수 있게 되었다.

- AI 윤리 책임감: 구글은 AI의 윤리적 사용과 보안성을 강화하기 위해 AI 기반 레드팀(Red Team) 및 신스ID(SynthID)와 같은 기술을 도입했다. 또한, AI의 투명성을 높이고 책임감 있는 GenAI를 개발하기 위한 다양한 툴킷을 공개하며, AI 기술의 안전한 활용을 위한 노력을 지속하고 있다.



출처: 구글코리아 블로그

[그림 2-1] Gemini 1.5와 타 모델 간 비교

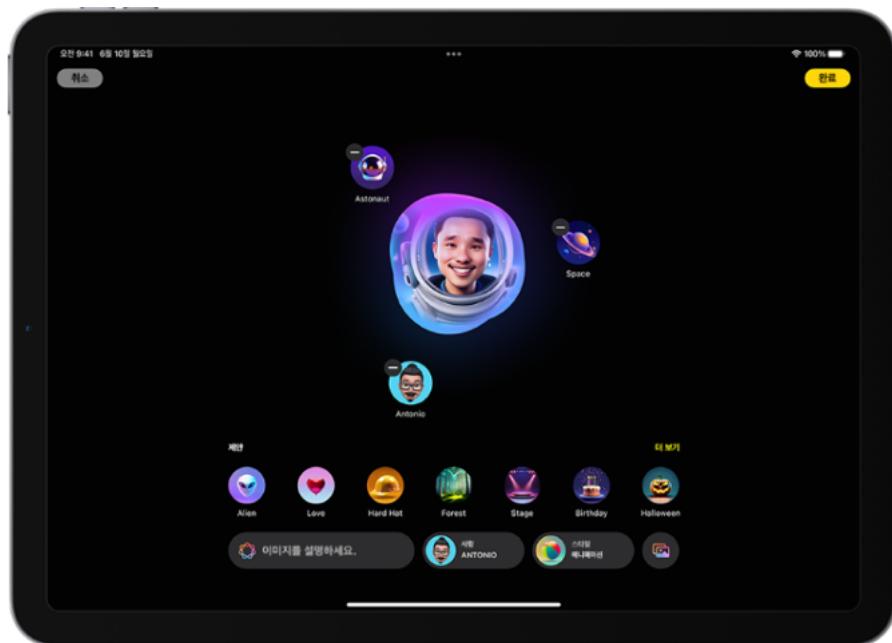
2024년 5월에 진행된 구글 I/O는 AI 기술이 단순한 도구를 넘어 창의적 작업, 검색, 교육 등 다양한 분야에 깊이 통합되고 있음을 보여주는 중요한 행사였다. 향후 구글의 AI 기술 발전이 더욱 가속화될 것으로 예상되며, 이에 대한 보다 구체적인 100가지 발표 내용은 [1]에서 확인할 수 있다. 구글 I/O는 AI 기술의 진화와 함께 일상생활 및 업무 환경에서의 변화를 예고하며, 구글이 AI 분야에서의 리더십을 확고히 하고 있음을 보여준다.



2.4.2. Apple

2024년 6월, 애플은 아이폰, 아이패드, 맥에 강력한 GenAI 모델을 탑재한 개인용 인공 지능 시스템인 Apple Intelligence를 공개했다. 이 시스템은 iOS 18, iPadOS 18, macOS Sequoia에 통합되어, Apple Silicon의 성능을 최대한 활용하여 사용자에게 개인화되고 맥락을 이해하는 AI 기능을 제공한다. Apple Intelligence는 언어 이해, 이미지 생성 및 인식 기술을 통해 일상적인 작업을 간소화하고, 사용자 경험을 혁신적으로 개선한다.

- 개인용 AI 시스템 탑재: Apple Intelligence는 Apple Silicon의 성능을 기반으로, 사용자에게 개인화된 AI 기능을 제공한다. 이 시스템은 언어 및 이미지 처리 기술을 통해 다양한 스마트 기능을 지원하며, 사용자의 일상적인 작업을 더욱 효율적으로 만든다. iOS 18, iPadOS 18, macOS Sequoia에 탑재되어, 사용자 기기에서 실시간으로 작동한다.
- 언어 이해 및 생성 기능: Apple Intelligence는 Mail, 메모, Pages 등 다양한 애플리케이션에서 텍스트 재작성, 교정, 요약 기능을 제공한다. 예를 들어, 이메일 내용을 요약하거나 맞춤형 이메일 작성을 지원하며, 문법 교정 기능도 강화되었다. 또한, Siri는 더욱 자연스러운 대화와 문맥 이해 능력을 갖추어 사용자의 요구를 빠르고 정확하게 처리한다. Image Playground와 같은 도구는 사용자가 텍스트 설명을 통해 다양한 스타일의 이미지를 생성할 수 있도록 지원한다.
- 비공개 클라우드 컴퓨팅 지원: Apple Intelligence는 사용자 데이터 보호를 최우선으로 하면서도 강력한 처리 능력을 제공하는 비공개 클라우드 컴퓨팅 시스템을 도입했다. 기본적으로는 온디바이스 처리를 중심으로 작동하지만, 복잡한 작업이 필요할 경우 Apple Silicon 기반 서버에서 클라우드를 활용해 처리한다. 이 시스템은 성능과 보안을 동시에 제공하며, 사용자의 동의 하에 클라우드와 통신하여 복잡한 작업을 처리한다.
- ChatGPT 기능 통합: 애플은 OpenAI의 ChatGPT를 iOS 18, iPadOS 18, macOS Sequoia에 통합했다. 이를 통해 Siri는 전문적인 정보를 제공하고, 사용자의 요청에 맞는 데이터를 검색 및 생성할 수 있다. ChatGPT와의 협업은 Apple 제품이 사용자의 개인적인 필요와 맥락을 더욱 정확히 반영하여, 직관적이고 효율적인 사용 경험을 제공할 것으로 기대된다.



출처: 애플 공식 블로그

[그림 2-2] Image Playground 예시

Apple Intelligence는 AI 기술을 사용자의 일상에 자연스럽게 통합하여, 더욱 스마트하고 개인화된 환경을 제공할 것으로 예상된다. 이번 발표는 애플이 AI 분야에서의 혁신을 통해 사용자 경험을 한 단계 업그레이드하려는 의지를 보여준다. 보다 구체적인 내용은 [2]에서 확인할 수 있다. Apple Intelligence의 도입은 AI 기술이 단순한 도구를 넘어 사용자의 삶 전반에 깊이 통합되는 새로운 시대를 열 것으로 기대된다.

2.4.3. OpenAI

2024년 12월, OpenAI는 '12 Days of OpenAI' 행사를 통해 12일 동안 AI 기술의 새로운 방향성을 제시하며 매일 주요 기능과 혁신을 공개했다. 이를 세 가지 관점으로 나누어 정리해보고자 한다.

- Day 1~4: AI 기술의 진보
 - Day 1: 고성능 추론 모델 'o1'을 공개하며, Pro 플랜 사용자들에게 향상된 추론 능력과 안정성을 제공했다. 특히, 자연어 처리 및 코드 해석 성능이 크게 개선되었으며, 보다 정교한 논리적 추론이 가능해졌다.
 - Day 2: 강화학습 기반 파인튜닝(Reinforcement Fine-Tuning, RFT) 기법을 소개하며, AI가 사고



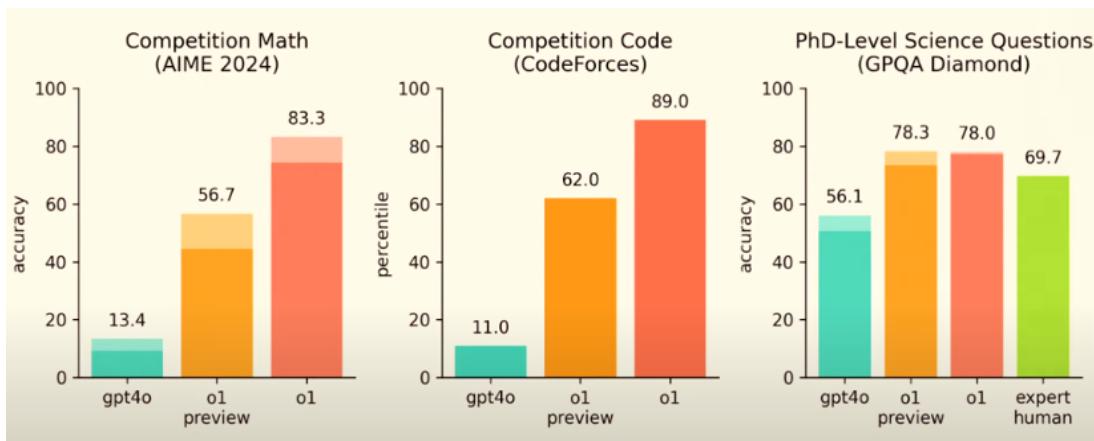
과정 자체를 최적화하여 전문 분야의 문제를 더욱 효과적으로 해결하도록 했다. 이를 통해 의료, 법률, 금융 등 고도화된 산업에서의 AI 활용 가능성이 확대되었다.

- Day 3: 텍스트 기반 고해상도 비디오 생성 모델 ‘Sora’를 발표하며, 창작의 범위를 확장했다. Sora는 스타일 지정, 장면 전환, 다양한 촬영 기법까지 지원하며, 단순한 애니메이션 제작을 넘어 영화 제작 수준의 결과물을 가능하게 한다.
- Day 4: 협업형 에디터 ‘Canvas’를 도입하여, 실시간 문서 편집과 Python 코드 실행을 지원함으로써 사용자 생산성을 높였다. 특히, 다중 사용자 협업 기능이 강화되어, 실시간 피드백과 코드 테스트가 보다 원활해졌다.
- Day 5~8: 사용자 경험 확장
 - Day 5: Apple 생태계와의 통합을 통해 iOS, iPadOS, MacOS에서 ChatGPT를 네이티브로 사용할 수 있도록 했다. 또한, Siri, 작문 도구, 카메라 제어 기능을 지원하여 사용성을 대폭 강화했다. Apple의 보안 기능과 연계하여 더욱 안전한 개인 맞춤형 AI 환경을 구축했다.
 - Day 6: 고급 음성 모드, 화면 공유, 그리고 ‘산타 모드’를 추가하여 멀티모달 상호작용과 사용자 참여도를 높였다. 음성 모드는 보다 자연스러운 대화 흐름을 지원하며, 감정 인식 기능이 추가되어 보다 인간적인 상호작용이 가능해졌다.
 - Day 7: ‘Projects’ 기능을 도입하여 작업을 파일 단위로 관리할 수 있도록 했으며, 맞춤형 지침을 제공하여 효율적인 협업 환경을 조성했다. 이를 통해 연구팀, 개발자, 콘텐츠 제작자들이 보다 체계적으로 프로젝트를 진행할 수 있게 되었다.
 - Day 8: 실시간 웹 검색 기능을 확장하여 모든 사용자에게 더 빠르고 접근성이 뛰어난 검색 경험을 제공했다. 검색 결과에 대한 요약 및 출처 제공 기능이 추가되었으며, 이를 통해 정보의 신뢰성을 높였다.
- Day 9~12: 개발자와 글로벌 접근성 강화
 - Day 9: 개발자 API 기능을 확장하여 GPT-4.1 및 새로운 도구를 선보이며, 보다 유연한 애플리케이션 통합을 가능하게 했다. 특히, API 요청 속도와 비용 최적화가 이루어져, 대규모 애플리케이션에서도 안정적인 성능을 유지할 수 있도록 개선되었다.
 - Day 10: 전화 및 WhatsApp 지원을 추가하여, 인터넷 연결이 제한된 환경에서도 AI에 접근할 수 있도록 개선했다. 이는 개발도상국이나 네트워크 인프라가 부족한 지역에서도 AI 서비스의 활용도를 높이는 계기가 되었다.
 - Day 11: 데스크톱 환경과의 통합을 통해 IDE(예: Xcode) 및 생산성 도구(예: Notion)와 연결하여 실시간 협업 및 자동화 기능을 강화했다. 특히, 코드 오류 감지 및 자동 수정 기능이 추가되어,



개발자들의 생산성을 크게 향상시켰다.

- Day 12: 최신 모델 'o3' 및 'o3-mini'를 발표하며, 더욱 향상된 성능과 확장성을 제공했다. 모델의 응답 속도와 효율성이 개선되었으며, 새로운 압축 기술을 도입해 동일한 하드웨어에서 더 높은 성능을 발휘할 수 있도록 최적화되었다.



출처: OpenAI

[그림 2-3] 다양한 업무에서의 모델 간 성능 비교

OpenAI는 AI 기술이 텍스트, 비디오, 음성, 코드, 문서 등 다양한 분야로 확장되고 있음을 보여주었다. 이는 사용자 경험, 창의성, 생산성을 극대화하며 AI의 실용적 가치를 넓히는 중요한 계기가 될 것으로 기대된다. 보다 구체적인 내용은 [3]에서 확인할 수 있다.

2.5. Agentic AI와 Physical AI

GenAI가 발전함에 따라 AI 에이전트가 중요한 응용 기술로 주목받고 있다. 2025년 1월 CES에서 엔비디아의 젠슨 황은 키노트 연설을 통해 이미 AI 기술이 Agentic AI 단계로 접어들었고, 곧 Physical AI로 발전할 것임을 언급하여 다음 단계의 기술 발전에 대한 기대감을 높였다. AI 에이전트 또는 Agentic AI라고도 표현되는 이 기술은 AI가 인간을 대행하여 자율적으로 업무를 수행하는 주체(Agency)가 될 수 있음을 의미한다. GenAI 이전에도 에이전트 기술이 있었으나, GenAI 이전과 이후의 큰 차이점은 자율적으로 판단해서 수행할 수 있는 능력이다. 이전 세대의 에이전트는 인간을 대신하여 업무를 수행하기 위해 사전에 업무의 범위가 구체적으로 정의되어 있어야 했으나, GenAI 기반의 에이전트는 주어진 업무를



수행하기 위해 필요한 단계를 스스로 판단하고 제어할 수 있다는 차이가 있다.

AI 에이전트의 주요 특징을 구글이 최근 발행한 백서 “Agents”를 통해 구체적으로 살펴보면 다음과 같다 [4]. 첫째, AI 에이전트는 언어모델에 기반한 추론(Reasoning)을 통해 할 일을 계획하고 의사결정을 내릴 수 있다. 이때 Chain-of-Thought과 같은 추론 프레임워크가 활용된다. 둘째, 에이전트는 언어모델 외부의 실시간 정보를 활용하고 외부 세계에서 쓸 수 있는 결과를 얻기 위해 도구들을 사용한다. 예를 들면, 에이전트가 예약을 수행하기 위해서는 외부의 웹사이트에서 제공하는 예약 API와 같은 연결고리가 필요하다. 셋째, 에이전트가 업무를 수행하기 위해서는 자신의 상태를 알고 업무의 진행 상황을 평가하고 업데이트할 수 있어야 하며, 장기 기억(Long-Term Memory)을 통해 관리할 수 있는 능력도 필요하다. 이러한 각 영역의 발전은 AI 에이전트가 수행할 수 있는 업무의 범위를 넓히고 정확도도 향상시킬 수 있을 것이다.

일반 사용자가 에이전트를 사용하는 방식은 2024년 1월, Rabbit이 발표한 Rabbit r1이라는 제품으로 엿볼 수 있다. 이 제품은 화면과 카메라, 마이크, 휠 등으로 구성된 손바닥만 한 소형 인터넷 기기로, Perplexity AI의 GenAI 기술을 활용, 음성 대화와 휠 조작 만으로 택시 호출이나 음식 배달 등을 진행할 수 있다고 홍보되었다. 실제 제품은 기대에 못 미치는 사용성을 보였으나, GenAI가 제공하는 대화형 에이전트 기반의 인터페이스는 앞으로 많은 서비스에서 기본적인 형태로 자리잡을 것이다. 특히, 사용자의 지시가 명확하지 않을 경우 에이전트가 질문을 통해 구체 내용을 파악할 수 있어 기존의 복잡한 사용자 인터페이스는 인간에게 더 직관적이고 친숙한 것으로 변화하게 될 것이다.

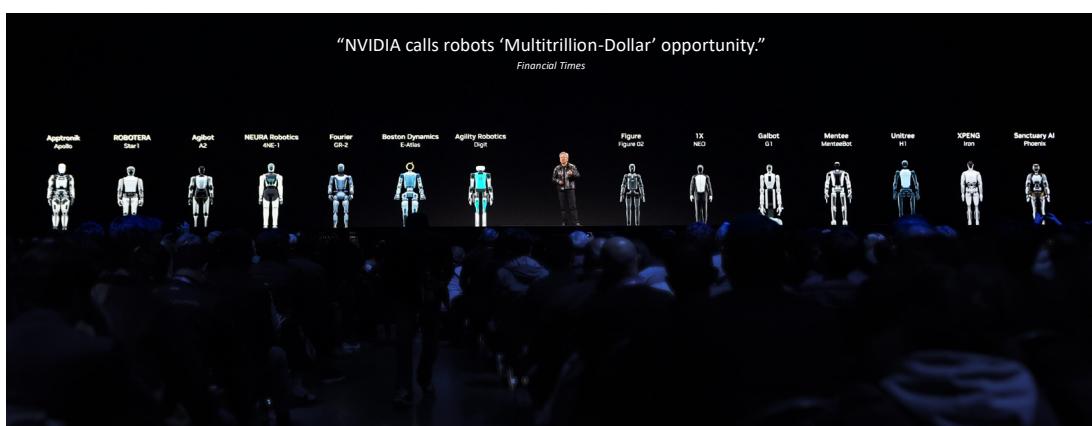
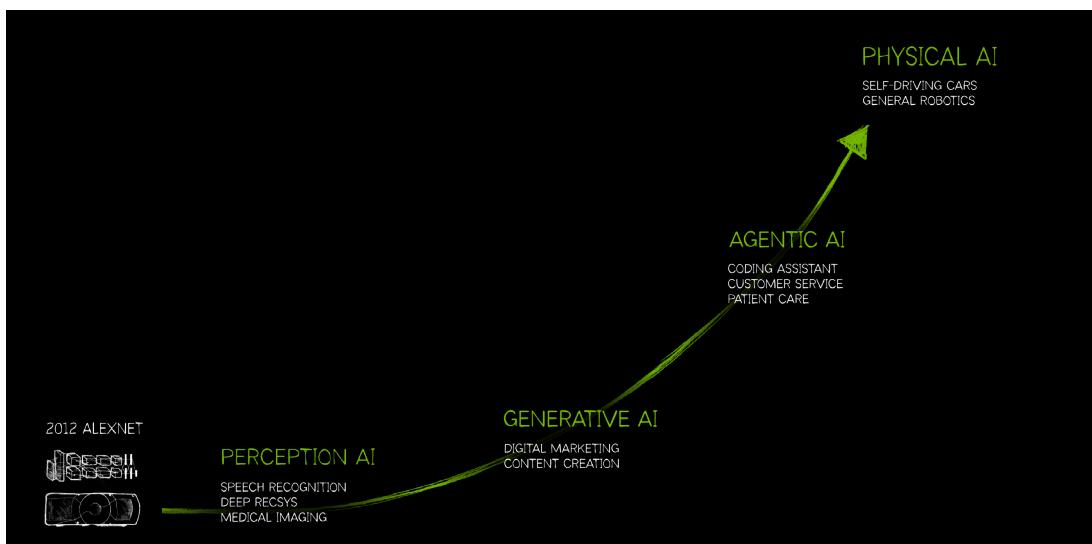
이러한 AI 에이전트는 앞으로 개인의 일정관리나 잡무를 처리하는 것은 물론, 복잡한 공장이나 네트워크 상의 이상 신호를 감지하고 이를 해결하기 위한 다음 행동을 자율적으로 결정할 수 있어, 단순히 인간을 보조하는 것을 넘어 다수의 AI 에이전트들이 공장이나 네트워크를 하나의 유기체와 같이 운영하는 것(Self-Organizing Network)도 가까운 미래에 가능할 것으로 보인다. 개별 에이전트들이 자율적으로 서로 소통하며 업무를 수행하는 미래를 엿볼 수 있는 연구 결과가 있다 [5]. 이 연구에서는 가상의 마을을 설정하고 그 안에 25명의 생성형 에이전트를 배치하여 시뮬레이션을 진행했다. 각 에이전트에는 환경을 인식하는 기능, 기억을 저장하는 구조, 행동을 되돌아보는 자기 성찰과 미래의 행동을 계획하는 능력을 부여했고, 목표를 지정하여 자율적으로 수행하게 하였다. 결과를 살펴보면, 각자 미션을 부여받은 에이전트들은 다른 에이전트들과 지속적으로 상호작용하고 상황을 인지하면서 목표를 달성하기 위한 단계들을 밟아나갔고, 그 과정에서 에이전트들과 사회 관계를 형성하며 창의적인 방식으로 문제를 해결해나가는 것을 볼 수 있었다. 이처럼 자율적으로 행동하는 에이전트는 상호작용하고 사회 관계를 형성하는 등 새로운 방식으로 문제를 해결할 수 있을 것으로 전망된다.

에이전트 기술이 더욱 발전하여 Physical AI 또는 Embodied AI 단계에 이르러 로봇 형태의 실체를 현실세계에 가지게 되면 그 서비스 영역은 인간 노동의 상당부분을 대체할 것이라고 볼 수 있다. 젠슨 황은



2. GenAI 서비스 및 기술 동향

CES 키노트에서 엔비디아와 협력하고 있는 휴머노이드 18종을 소개하며 휴머노이드가 인간의 곁에서 다양한 업무를 수행하는 때가 머지 않았음을 시사했다. 최근까지 쉽지 않을 것으로 판단했던 로봇의 학습은 현실 또는 가상 공간에서 동작을 보고 따라 하는 모방 학습(Imitation Learning) 기술이 등장하면서 빠르게 발전하고 있다. 앞서 기술한 AI 에이전트 기술과 현실 세계를 인식하고 제어하는 로봇 기술이 성숙하여 결합될 때 Physical AI가 현실화되고 서비스로서의 로봇이 보편적으로 자리를 잡게 될 것이다.



〈Nvidia의 CES 2025 키노트 자료 중〉

3. AI-RAN 서비스 및 기술 동향





3. AI-RAN 서비스 및 기술 동향

3.1. AI-RAN 필요성

3.1.1. 비즈니스 관점

통신 산업은 빠르게 발전하는 기술과 증가하는 데이터 수요로 인해 상당한 자본 지출 압박에 직면해 있다. GSMA(Global System for Mobile Communications Association)의 The Mobile Economy 2024 보고서에 따르면 글로벌 모바일 시장에서 통신사 총수익은 2023년 1조 1,100억 달러에서 2030년 1조 2,500억 달러로 성장할 것으로 예상되며, 이는 1.74%의 완만한 연평균 성장률을 나타낸다. 그러나 2030년까지 총 자본 투자는 1조 5,000억 달러로 추산되며, 이는 총 단일 연도 수익을 초과하여 적자를 예상하는 수치이다. 이는 전 세계 통신사가 직면한 중대한 과제로 이를 극복하기 위한 대안으로 AI-RAN이 검토되고 있다. 통신사의 비즈니스에 AI-RAN을 채택함으로써 네트워크 투자 대비 이득을 극대화하고, 새로운 비즈니스 모델 발굴을 통해 지속 가능한 성장을 유도할 수 있을 것으로 기대된다.

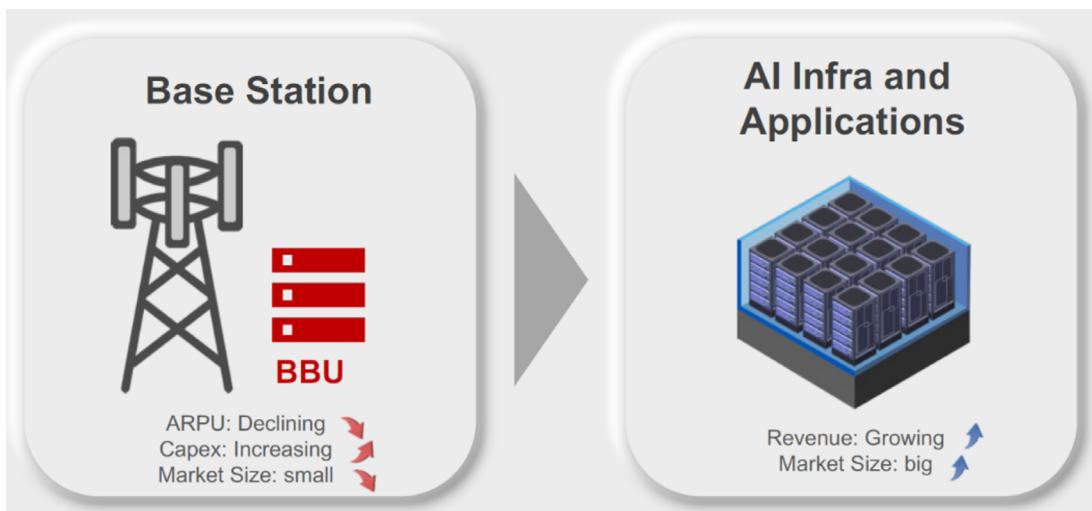


출처: GSMA, The Mobile Economy 2024

[그림 3-1] 통신사의 총수익 변화



3. AI-RAN 서비스 및 기술 동향



출처: AI-RAN Alliance

[그림 3-2] AI-RAN을 활용한 Telco의 비즈니스 변화

3.1.2. 기술 관점

2장에서 설명된 다양한 GenAI 서비스를 수용하기 위한 차세대 이동통신 기술의 발전은 무선 액세스 네트워크(Radio Access Network, RAN)의 구조와 운영에 있어 복잡성을 가중시키고 있다. 기존의 수작업 기반 네트워크 운용 방식으로는, 다양한 GenAI 서비스 제공 시 예상되는 급변하는 트래픽 패턴과 다양한 서비스 요구사항에 효과적으로 대응하기 어렵다. 특히, 거대 다중 입력 다중 출력(Massive MIMO), 빔포밍(Beamforming), 네트워크 슬라이싱(Network Slicing) 등의 기술은 RAN의 실시간 자원 운영을 더욱 어렵게 만들고 있다.

이에 따라 네트워크 운용의 자동화 및 지능화를 위한 새로운 패러다임이 요구되며, 이에 따라 이를 극복하기 위한 AI 기반의 RAN, 즉 AI-RAN의 필요성이 점점 높아지고 있다. AI-RAN은 단순한 자동화를 넘어, 네트워크가 스스로 판단하고 진화할 수 있는 구조를 지향하며, 특히 6G 시대에는 사용자 중심의 초지능형 네트워크가 요구되므로, 이를 위한 기술적 도약으로 AI-RAN 기술은 이동통신시스템에서 다양한 GenAI 서비스를 제공하기 위한 필수적인 기술이 될 것으로 예상된다.

[표 3-1] 기존 RAN의 기술적 한계

네트워크 복잡도의
증가

다중 셀 환경, 사용자 수의 폭발적 증가, 서비스 분류의 다양화는 RAN 구조의 복잡성을 증가시키고 있다. 특히, 빔관리(Beam Management), 간섭관리(Interference Management) 등에서 실시간 최적화가 어려운 문제가 존재한다.

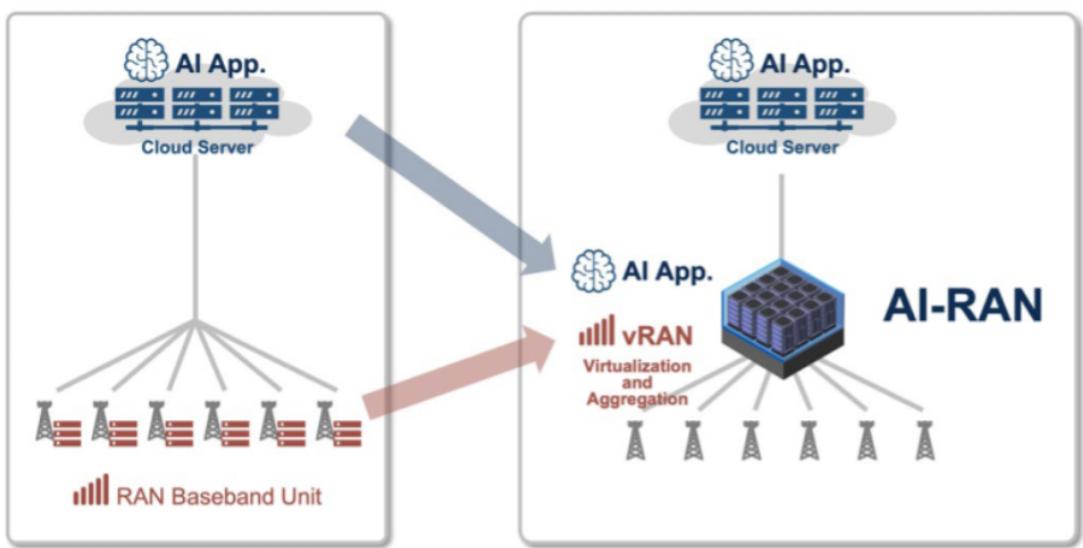


운용 및 유지관리의 비효율성	현재 대부분의 네트워크 운용은 전문가에 의한 수동 설정과 규칙 기반 알고리즘에 의존하고 있으며, 이는 네트워크 구성 변화에 신속히 대응하기 어렵다.
트래픽 및 자원 예측 의 부정확성	시간, 장소, 사용자 패턴에 따라 급변하는 트래픽 특성은 정적인 운용 정책으로는 수용이 어렵다. 이것은 자원 낭비, QoS 저하, 에너지 비효율이라는 결과로 이어질 수 있다.

3.2. AI-RAN 개요

AI-RAN(Artificial Intelligence in Radio Access Network)은 모바일 통신 시스템의 RAN에 AI 기술을 통합하는 것을 말한다. 이러한 접근 방식은 복잡한 작업을 자동화하고, 리소스 활용을 최적화하고, 예측 분석 및 동적 네트워크 관리와 같은 고급 기능을 활성화하여 RAN의 성능, 효율성 및 적응성을 향상시킬 수 있다.

RAN은 모바일 네트워크의 중요한 구성 요소로써, 무선 액세스를 통해 사용자 장치(예: 스마트폰, IoT 장치)를 코어 네트워크에 연결시키는 중요한 역할을 한다. 여기에 인공지능을 통합함으로써 RAN이 더욱 지능화되어 5G, 6G를 넘어서는 미래의 무선 네트워크의 요구 사항도 충족시킬 수 있는 기술로 기대되고 있다.



출처: SoftBank

[그림 3-3] AI 와 RAN의 통합



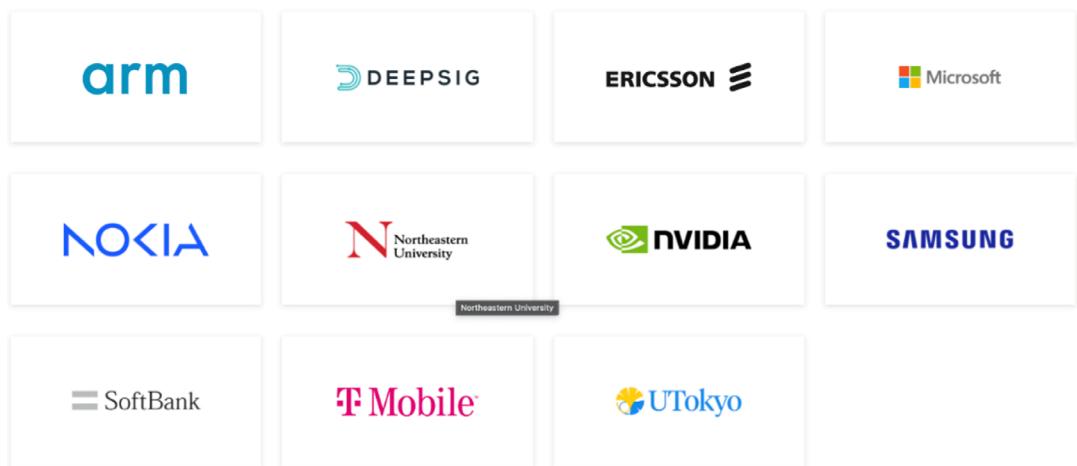
3. AI-RAN 서비스 및 기술 동향

이러한 AI-RAN을 활성화하기 위해 '24.3월 MWC(Mobile World Congress) 2024 기간 중 미국의 NVIDIA와 일본의 SoftBank 주도로 AI-RAN Alliance가 창립되었으며, 현재 AI-for-RAN, AI-and-RAN 및 AI-on-RAN 등 3개의 Working Group(WG)과 Technical Steering Committee(TSC) 및 Marketing Steering Committee(MSC)의 세부 그룹으로 운영되고 있다.

AI-RAN Alliance는 비표준화기구를 지향하고 있어 별도의 표준화 활동은 하지 않으며, AI-RAN 관련 기술의 활성화를 위해, 관련 기술을 보유한 회원사들을 결집하고, AI-for-RAN, AI-and-RAN 및 AI-on-RAN 관련 기술 및 서비스 개발을 지원하기 위한 기술개발 절차 및 데이터 포맷을 개발하고, 회원사들이 개발한 AI-RAN 기술을 검증하기 위한 실험실 구축을 목표로 하고 있다. 또한, MWC 등 유수의 전시회에 회원사들이 개발한 결과물을 AI-RAN Alliance 이름으로 전시함으로써 AI-RAN Alliance의 결과를 홍보할 계획이다.

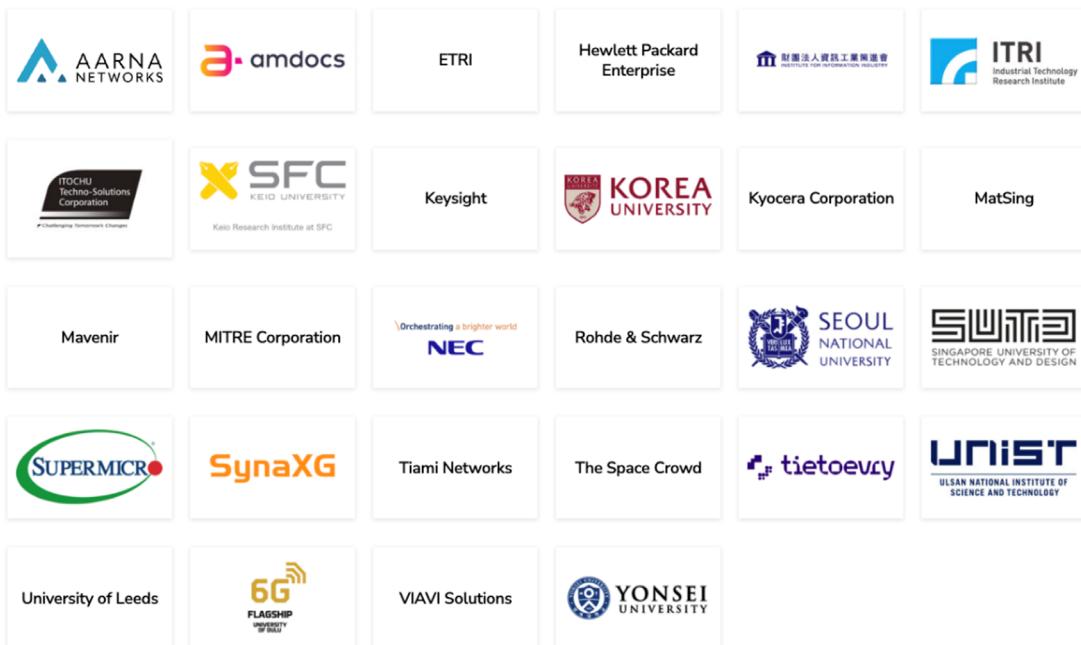
AI-RAN Alliance는 온라인 회의를 통해 회원사들의 진행상황을 수시로 교류하고, 매년 두번의 현장 회의(5월~6월중 1차 오프라인 회의, 11월중 2차 오프라인 회의)를 통해 Work Item별 기술개발 현황, 결과물 점검 및 MWC 등 전시회를 준비한다. 이번 '24.11월 오프라인 미팅에서는 MWC 2025에 데모하기 위한 기술 및 방법론이 논의되었다

AI-RAN Alliance의 '24. 11월 당시의 멤버 구성은 다음 그림과 같다.



출처: AI-RAN Alliance

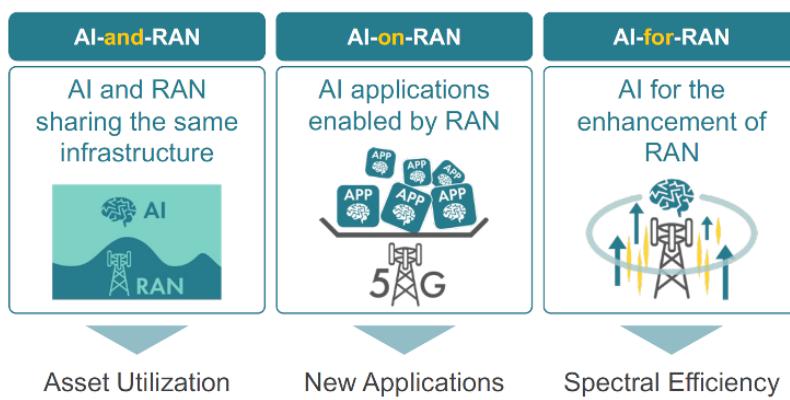
[그림 3-4] AI-RAN Alliance의 Founding member



출처: AI-RAN Alliance

[그림 3-5] AI-RAN Alliance의 General member

AI-RAN은 차세대 RAN를 위한 인공지능(AI) 통합 프레임워크로, 네트워크 인프라의 효율성을 극대화하고 네트워크의 성능을 향상시키기 위해 도입되었다. AI-RAN Alliance의 6G Generative AI-RAN은 AI와 RAN의 결합 방식을 AI-and-RAN, AI-on-RAN, AI-for-RAN의 세 가지 방식으로 분류하고 있으며, 각 결합 방식의 특성과 역할은 다음과 같다.



출처: AI-RAN Alliance

[그림 3-6] AI와 RAN의 결합방식 분류



■ AI-and-RAN

AI-and-RAN 접근 방식은 AI와 RAN이 동일한 물리적 인프라를 공유하며 협력하는 방식을 의미한다. 이 방식은 AI와 네트워크 인프라 간의 밀접한 통합을 통해 데이터 처리와 네트워크 관리의 효율성을 높인다. AI와 RAN이 하나의 물리적 네트워크 인프라 내에서 작동함으로써, 자원 활용이 최적화되고, 데이터 흐름이 보다 원활하게 이루어질 수 있다.

■ AI-on-RAN

AI-on-RAN 접근 방식은 RAN이 AI 응용 프로그램(application)을 지원하는 역할을 한다. 이 접근 방식은 특히 5G 및 그 이후의 미래 네트워크 환경에서 효과적일 수 있다. RAN은 다양한 AI 응용 프로그램의 플랫폼 역할을 하며, 이로 인해 실시간 데이터 분석, 사용자 요구 예측, 네트워크 리소스 최적화와 같은 AI 응용 프로그램이 좀 더 원활하게 실행될 수 있다. 이를 통해 RAN은 AI 기술의 기능을 더욱 강화할 수 있어서, 보다 지능적이고 응답성이 뛰어난 네트워크를 구현할 수 있다.

■ AI-for-RAN

AI-for-RAN 접근 방식은 RAN의 성능 향상을 위해 AI를 적용하는 방식이다. 이 접근 방식은 AI가 RAN의 자원 할당, 네트워크 상태 분석, 네트워크 성능 최적화와 같은 작업에 직접적으로 기여하여, 실시간 네트워크 인프라의 데이터 분석을 통해 RAN의 효율성을 극대화하고 네트워크 서비스 품질을 높이는 것을 목표로 한다. AI의 예측 및 분석 기능을 통해 RAN은 자원을 효율적으로 관리하고, 트래픽 혼잡을 방지하며, 사용자 경험을 향상시킬 수 있다.

AI-RAN은 AI와 RAN의 결합을 통해 무선 네트워크의 효율성과 성능을 극대화하는 데 중점을 두고 있다. AI-and-RAN, AI-on-RAN, AI-for-RAN의 세 가지 결합 방식을 통해 다양한 네트워크 환경에서 AI와 RAN의 협력이 가능해지며, 이를 통해 차세대 무선 네트워크에서 보다 지능적이고 효율적인 서비스 제공이 가능하게 된다. 특히 GenAI를 활용할 경우, 사용자 요구 예측, 네트워크 자원 최적화, 장애 사전 대응 등 고도화된 기능을 통해 AI-RAN은 미래 네트워크 환경에서 핵심적인 인프라로 자리잡을 것으로 기대된다.

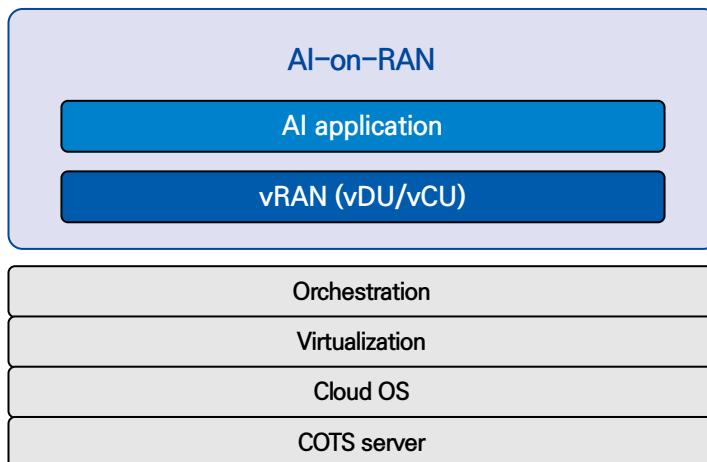
본 백서에서는 AI-and-RAN을 제외한 AI-on-RAN과 AI-for-RAN 관련 연구사례에 대해 설명한다.



3.3. AI-on-RAN

AI-RAN Alliance의 WG3에서 연구되고 있으며 현재 및 미래 시스템에 대한 핵심 RAN 요구 사항을 파악하여 RAN 연결 및 인프라 연결을 통해 AI 응용 프로그램을 제공하고 벤치마킹하는 데 목표를 두고 있다. 이 그룹의 목표에는 현재 인공지능 중심 기술을 검토하고, 과제를 식별하고, 사용 사례를 정의하고, 테스트 계획을 개발하고, 시스템 구현 청사진을 제공하는 것이 포함되어 있다. AI-RAN Alliance WG3 작업에서는 AI 응용 프로그램의 QoE 향상시키는 데 필수적으로 필요한 차별화된 네트워크 연결 및 디바이스(device) 기능을 제공하는 것을 중요시하고 있다.

AI-on-RAN의 구조는 [그림 3-7]과 같으며 하위 COTS server, Cloud OS, Virtualization, Orchestration의 기능을 갖는 동일한 AI-RAN 플랫폼에서 vRAN 과 AI 응용프로그램이 동시에 탑재되어 수행되는 구조를 갖고 있다. 이는 종래 MEC에서 동작하던 AI 응용 프로그램이 RAN 내부로 수용되는 효과를 가질 수 있다. 본 AI 응용 프로그램은 엣지(Edge) AI 응용 프로그램으로도 불린다.



[그림 3-7] AI-on-RAN의 구조

AI-on-RAN은 다음과 같은 분야에서 중점적으로 연구되고 있다.

- AI 기반 멀티미디어 응용 프로그램: 비디오 분석, 증강/가상현실(AR/VR), 초실감 게임, 텍타일 인터넷, 훌로그램 통신 등의 AI 기반 멀티미디어 응용 프로그램 분야.
- AI 기반 크리티컬(Critical) 응용 프로그램: 침입탐지와 같은 AI 기반 보안 응용 프로그램, 그리고 헬스케어와 같은 크리티컬 응용 프로그램 분야.



3. AI-RAN 서비스 및 기술 동향

- AI 기반 자동화 및 제조 응용 프로그램: 제조 자동화를 위한 AI 기반 제조 응용 프로그램, 그리고 무인항공기(UAVs), 드론 및 무인 운반차(Automated Guided Vehicles, AGVs)를 위한 AI 기반 자율주행 응용 프로그램 분야.
- AI/GenAI 기반 네트워크 서비스: 고객중심 서비스, 사용자와의 상호작용, 위치관리 서비스를 포함하여 통신의 효율성 및 사용자와 응용 프로그램 사이의 중단 없는 협력을 가능하게 하는 AI(또는 GenAI) 기반 네트워크 기반 서비스 분야.
- 효율적인 AI/ML 모델 분할: 무선 링크와 디바이스의 배터리 상황을 고려하여 단말 디바이스와 RAN에 대한 AI/ML 모델을 분할하는 분야.

3.3.1. AI/ML 모델 분할 기반 스펙트럼 센싱

싱가폴의 SUTD(Singapore University of Technology and Design) 대학과 Keysight Technologies는 개인정보보호를 우선시하는 프라이버시 중심의 이미지 처리 및 적응형 AI/ML(AI/Machine Learning) 모델 분할을 위한 연구를 공동으로 수행하였다. 본 연구에서는 시간적으로 변화하는 동적 5G 통신환경에서 개인의 프라이버시 보호, 종단간(end-to-end) 지연 시간, 에너지 효율성 및 트래픽 처리량과 같은 중요한 지표가 어떻게 최적화될 수 있는지 보여주었다. 또한, 간섭, 재머(Jammer) 및 단말 이동성과 같은 다양한 어려운 조건에서 이미지 처리를 5G 트래픽 처리량을 고려하면서 Deep ML 모델을 적응적으로 분할하여 적용하였다.

동적으로 변화하는 5G 네트워크의 무선 환경에서 개인정보보호 중심 이미지 처리를 고정된 ML 모델 분할 형태로 처리할 경우 비효율성이 높아진다. 따라서 다음과 같은 문제들의 해결을 통해 효율성을 높이고자 하였다.

- 동적 무선 채널 조건:

간섭, 재머, 단말 이동성과 같은 요소로 인해 네트워크 환경은 지속적으로 변동되며, 이러한 조건은 ML 모델이 고정된 지점에서 분할될 경우 성능 저하를 일으킬 수 있다. 특히 이러한 동적인 환경은 실시간 AI 응용 프로그램에서 중요한 트래픽 처리량, 지연 시간 및 신뢰성 등에 직접적인 영향을 미친다.

- 성능 지표의 상충:

ML 모델의 고정 분할은 여러 성능 지표 간의 상충 문제를 야기할 수 있다. 우선, 지연 시간 측면에서는 열악한 네트워크 조건 하에서 더 높은 종단간 지연이 발생할 수 있다. 에너지 소비 측면에서는 성능 최적화 없이 모델이 작동할 경우 사용자 단말이 필요 이상으로 많은 에너지를 소모하게 되어 배터리 수명이 단축된다. 또한 개인 정보 보호 관점에서는, 고정된 분할 지점에서 보다 민감한 중간 데이터가 외부에 노출될 가능성이 있어 개인 정보 유출 위험이 커질 수 있다.



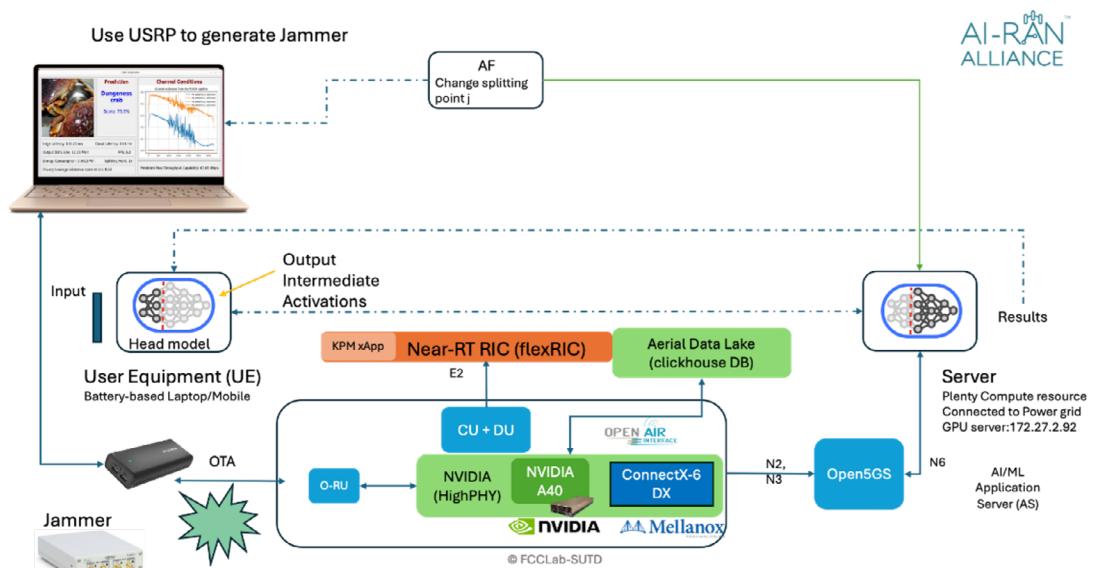
- 5G 네트워크의 비효율적 리소스 활용:

5G 환경에서는 네트워크 리소스를 효율적으로 활용하는 것이 필수적이며, 이는 다양한 트래픽 처리량에 대응하고 엣지와 클라우드 간의 처리 균형을 보장하는 데 중요한 요소가 된다. 그러나 고정 분할 방식은 이러한 자원 활용에 제약을 가할 수 있다.

- 적응성 부족:

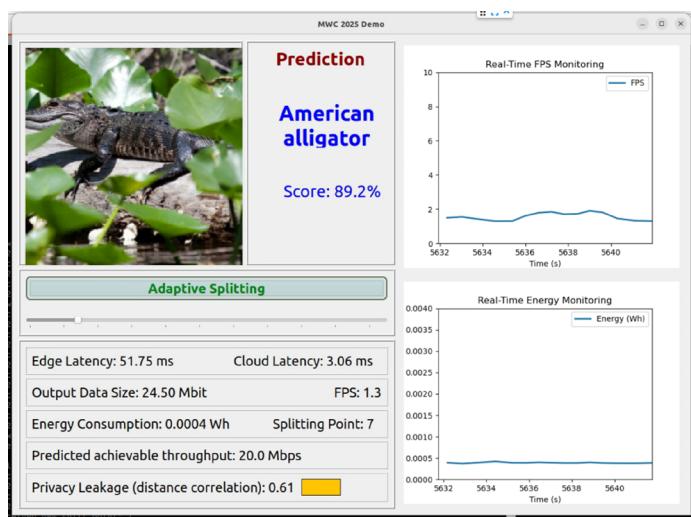
ML 모델이 고정된 방식으로 분할될 경우, 동적으로 변화하는 채널 조건에 효과적으로 적응하지 못하게 된다. 이는 실제 배포된 ML 모델에서의 성능 저하, 운영 비용의 증가, 그리고 전반적인 처리 비효율성으로 이어질 수 있다.

따라서, 실시간 AI 기반 스펙트럼 센싱에 적응형 ML 모델 분할을 도입함으로써, 이러한 문제를 효과적으로 해결하고 개인 프라이버시, 지역 시간, 에너지 소비 및 트래픽 처리량을 동시에 최적화하는 방법을 보여주었다. 본 연구는 매우 가변적인 통신환경에서도 개인 프라이버시 보호를 중요시하는 AI 응용 프로그램의 성능 향상에 도움을 줄 것으로 예상된다.



출처: AI-RAN Alliance

[그림 3-8] 동적 AI/ML 모델 분할기반 스펙트럼 센싱 구조



출처: AI-RAN Alliance

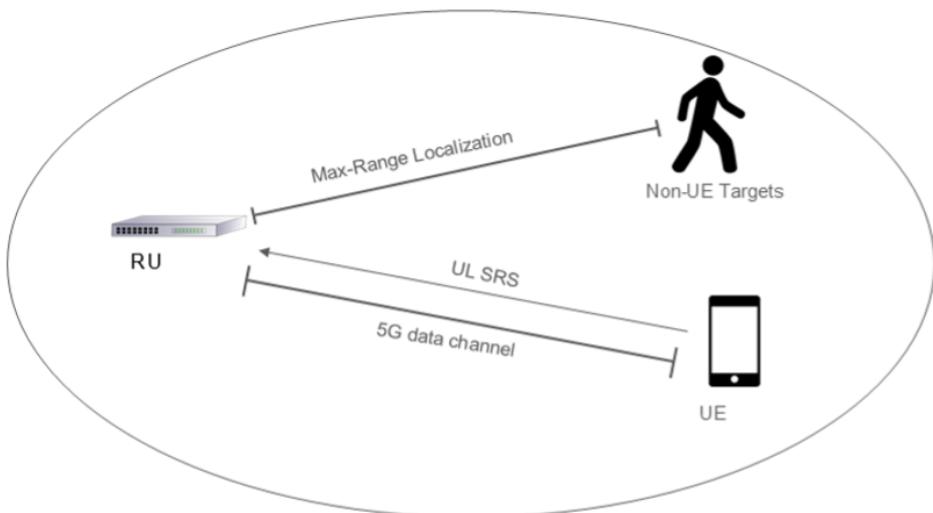
[그림 3-9] 동적 AI/ML 모델 분할기반 스펙트럼 센싱의 단말 GUI 화면

3.3.2. AI 기반 통신 및 센싱 결합

본 절에서는 AI와 결합된 기존 상용 5G 통신 인프라가 단말에게 통신 서비스를 제공하는 동시에, 표준 5G 무선신호를 활용해 보행자와 같은 비 단말 대상을 탐지하고 추적하는 기능을 어떻게 구현할 수 있는지를 설명한다. 특히, 5G 통신 네트워크는 단말의 상향링크(Uplink, UL) SRS(Sounding Reference Signal)의 반사신호를 분석하여 보행자와 같은 비 단말 대상을 감지한다. 종래에는 기지국의 DL(Downlink) PBCH(Physical Broadcast Channel)의 반사신호를 분석하여 단말 측에서 감지를 수행했지만, 본 연구에서는 단말에서 전송한 UL SRS의 반사신호를 분석하여 기지국 측에서 감지를 수행한다.

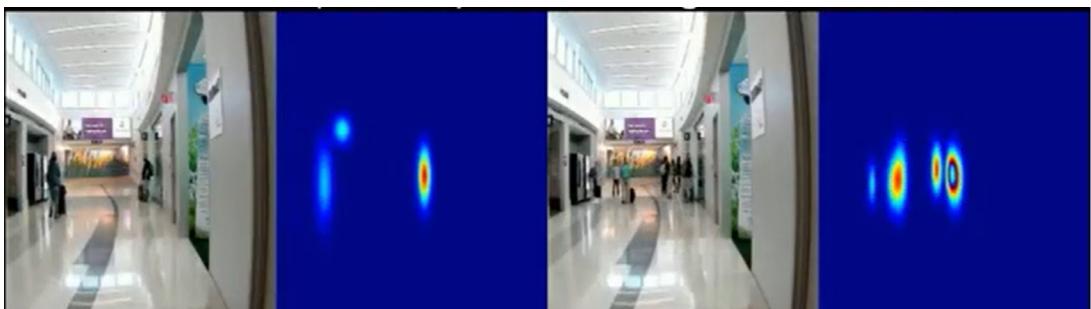
표준 5G 무선신호를 사용하여 특정 지역에서 단말이 아닌 보행자와 같은 대상의 존재를 감지하고 추적하기 위해서는 5G DL PBCH 및 UL SRS 신호가 감지기능에 활용될 수 있는지에 대한 연구가 필수적이며, 또한 상용화를 위해서는 현재 하드웨어 및 기능을 변경하지 않고 이를 달성하는 것이 매우 중요한 목표가 될 수 있다.

상용 5G 단말에서 전송한 UL SRS 신호는 기지국의 라디오 유닛(Radio Unit, RU)에서 수신되어 디지털 유닛(Digital Unit, DU)에서 처리되고, 여기서 High-PHY에서 추정된 채널 상태 정보(Channel State Information, CSI)는 RAN에 탑재된 추론 머신 러닝 모델에 제공된다. 최대 100MHz 대역폭을 사용하는 상용 중간 대역 5G RAN을 활용하였으며 상용 단말의 통신과 비 단말 대상 탐지 및 추적 기능은 동일한 무선 주파수 자원을 사용하였다.



출처: AI-RAN Alliance

[그림 3-10] AI 기반 통신 및 센싱 결합 시나리오



출처: AI-RAN Alliance

[그림 3-11] AI 기반 통신 및 센싱 결합 결과

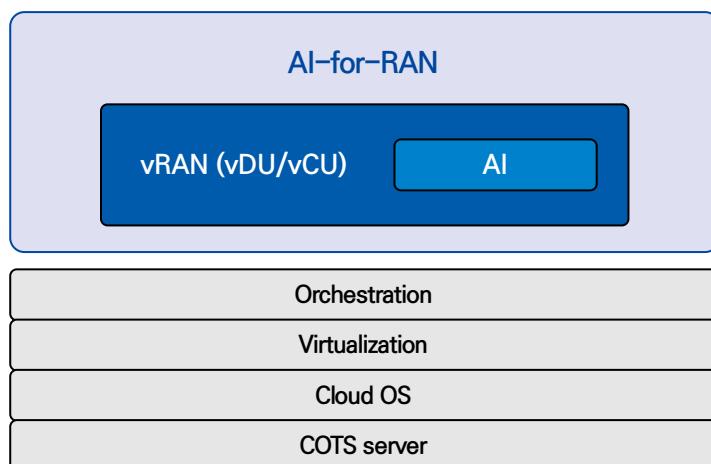
[그림 3-10]과 [그림 3-11]은 5G 통신과 동시에 보행자 탐지 및 추적 기능을 수행하는 시나리오와 히트맵 화면을 통해 보행자를 감지한 결과를 보여준다. 본 실험에서 사용된 5G 단말은 통신 성능 저하 없이, 미리 정해진 구역 내 보행자 수를 성공적으로 추정할 수 있었으며, 탐지 및 추적 결과는 대시보드 플랫폼을 통해 시연되었고, 사용자 단말의 원활한 통신은 전화 화면 시연을 통해 확인되었다. 이는 5G 단말의 UL SRS 신호를 활용하여 주변 환경을 감지하고 대상을 탐지함으로써, 5G 네트워크 지능화를 실현하는 데 있어 AI가 수행할 수 있는 역할 중 하나를 보여주는 사례이다.



3.4. AI-for-RAN

AI-RAN Alliance의 WG1에서 연구되고 있으며, AI를 이용하여 RAN 시스템의 효율성, 용량, 그리고 성능지표를 향상시키는 데 초점을 맞추고 있다. 또한, 본 그룹의 목표에는 문헌 검토 수행, 사용 사례 정의, 개념 증명(Proof-of-Concept)을 위한 시스템 구현이 포함되며, 특히 AI/ML의 활용을 통해 AI-RAN 산업을 선도하고 발전시키는 데 중점을 두고 있다.

AI-for-RAN의 구조는 [그림 3-12]와 같으며 하위 COTS server, Cloud OS, Virtualization, Orchestration의 기능을 갖는 물리적으로 동일한 AI-RAN 플랫폼에서 AI 기능이 vRAN 기능들과 밀접하게 연결되어 수행되는 구조를 갖고 있다.



[그림 3-12] AI-for-RAN 구조

AI-for-RAN은 다음과 같은 분야에서 중점적으로 연구되고 있다.

- AI 기반 무선 인터페이스 및 신호처리: RAN 시스템의 성능 및 효율성 향상과 오버헤드 감소를 위해 AI/ML 기술을 통해 기존 신호처리 방식을 대체하거나 증강할 수 있는 분야. 수신기 최적화, 학습기반 종단간 무선 인터페이스, 통신 및 센싱 결합(JCAS/ISAC), 분산형 다중 입력 다중 출력 (D-MIMO), 초거대 다중 입력 다중 출력(XL-MIMO) 등.
- 위치 및 빔 관리: 위치관리(Positioning)의 정확도를 향상시키고 빔포밍(Beamforming)의 민첩성과 성능을 향상시킬 수 있는 분야. SRS 와 PRS(Positioning Reference Signal) 기반 위치관리와 빔관리를 위한 CSI 압축 및 예측 등.



- 무선자원관리 및 스케줄링(scheduling): AI를 활용하여 무선 주파수 자원의 동적 할당과 사용률을 향상시킬 수 있는 분야. 트래픽 패턴 및 네트워크 환경 예측 기반 스케줄링. RAN 시스템 전체 성능 및 자원 최적화.
- 에너지 및 스펙트럼 효율성: RAN 시스템의 동작 요소들과 주파수 사용 최적화를 통해 RAN 시스템 전체의 소비 에너지를 감소시킬 수 있는 분야. 셀 에너지 절약 및 스펙트럼 센싱 및 공유 등.
- 네트워크 최적화 및 이상 징후 탐지: AI를 활용하여 네트워크의 이상 징후 예측과 검출, 그리고 네트워크 운영상의 위험요소를 감소시키고 네트워크 생존성을 향상시키는 분야. RAN 디지털 트윈 및 기지국 설치 장소 별 최적화, 이상 징후 탐지(Anomaly Detection), 장애 예측 및 완화, 무선 인터페이스 품질 보장, 시맨틱 통신(Semantic Communications) 등.

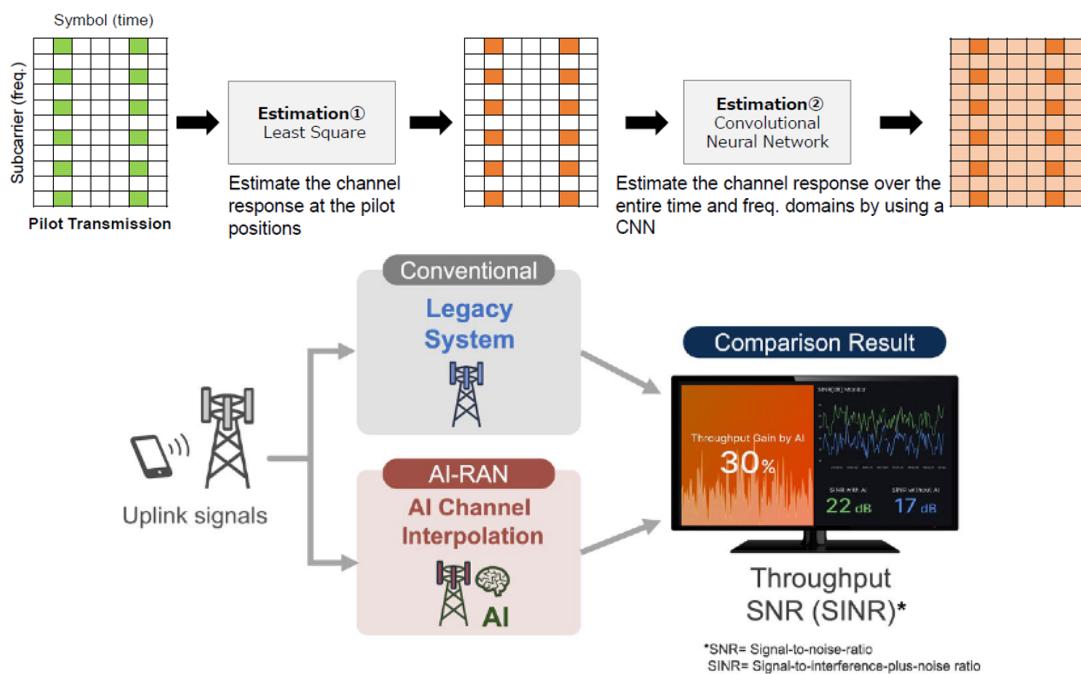
3.4.1. AI 기반 상향링크 채널 보간법

여러 기지국과 단말기가 있는 밀집 환경에서 무선 신호는 다중 경로 페이딩(multipath fading)으로 인해 왜곡될 수 있다. 이러한 환경에서 기존 신호 처리 기술은 무선 특성을 정확하게 추정하지 못해 트래픽 처리량(throughput)이 낮아지는 문제가 있었다.

미국의 NVIDIA와 일본의 SoftBank는 이러한 문제를 해결하기 위해 원래 이미지 분석에 사용되었던 CNN(Convolution Neural Network) 기반 AI-native 초고해상도 영상기술을 무선 신호 처리에 적용하여 저하된 신호를 AI 기반 보간법(Interpolation)을 통해 재구성한 후 상향링크 트래픽 처리량 개선에 적용하는 연구를 수행하였다. 실제 환경 조건에 기반하여 시뮬레이션을 통해 얻어진 무선 신호 데이터로 CNN 기반 AI 모델을 학습하고 상향링크 신호로 테스트하여 기존 신호 처리 기술에 비해 상향링크 처리량이 30% 향상되는 것을 시연하였다.



3. AI-RAN 서비스 및 기술 동향



출처: SoftBank

[그림 3-13] AI 기반 UL Channel Interpolation

3.4.2. AI 기반 PUSCH 채널 추정 기법

본 연구에서는 일반적으로 주파수 대역에서 주로 잡음 제거를 수행하는 AI 기반 채널 추정을 PUSCH(Physical Uplink Shared Channel) 수신 성능을 개선하는 데 적용하여 셀의 커버리지 확장에 중점을 두었다. PHY 계층에 AI를 적용하여 RAN의 상향링크 데이터 수신 성능 향상과 더불어 RAN의 기존 성능지표를 더욱 향상시키는 방안을 보여주고자 하였다.

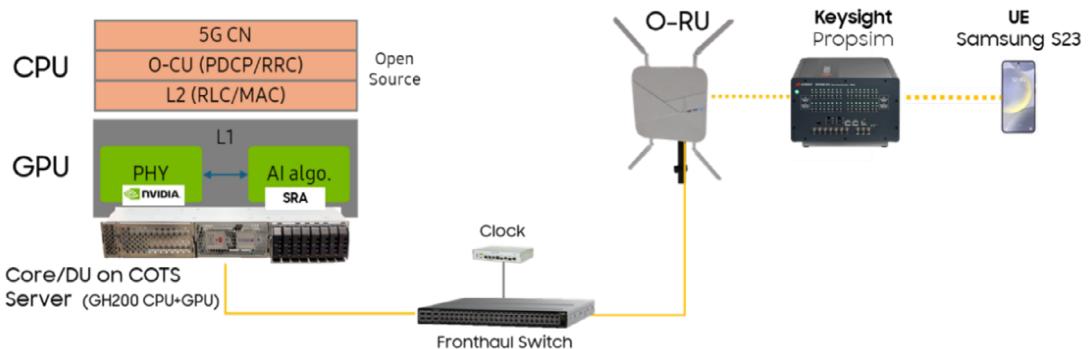
상용 5G 네트워크 운영 시 상향링크 성능에 의해 종종 커버리지와 용량이 제한되는 측면이 있다. 가장 큰 문제는 셀 엣지에서 단말이 특정 피크 전력 수준으로만 전송할 수 있어 기지국의 수신 SNR이 제한된다는 것이다. 본 연구에서는 상향링크 DMRS(DeModulation Reference Signal) 채널 추정 성능이 PUSCH 성능 열화에 중요한 원인 중 하나인 것을 보여주었다.

기존 상향링크 채널 추정 알고리즘은 일반적으로 과도한 노이즈로 인해 낮은 SNR 영역에서 원활한 성능을 보여주는 데 한계가 있었다. 따라서 본 연구에서는 상향링크 채널 추정에 AI를 새로운 도구로 사용하여 상향링크 MAC 트래픽 처리량을 개선할 수 있도록 상향링크 채널 추정 알고리즘을 재설계하여 커버리지 및 용량이 증가될 수 있도록 하였다.



연합 주파수-시간-공간 도메인에서 고차원 무선 채널 데이터를 처리하기 위해 AI를 활용하였고, 더 나은 채널 추정 결과를 얻기 위해 잡음이 있는 수신 DMRS 신호의 잡음을 제거하는 신경망 모델을 개발하였다. 다양한 3GPP 채널과 RF 손상 모델과 함께 다양한 사이트 데이터에 대한 레이 트레이싱 결과를 사용하여 AI 모델을 훈련하였고, 높은 SNR 필드 데이터와 서로 다른 SNR 및 PRB(Physical Resource Block) 크기를 적용하여 보강하였다. 결과적으로 본 연구의 실시간 AI 기반 채널 추정 알고리즘이 종래 채널 추정 알고리즘에 비해 트래픽 처리량이 개선되고 적용 범위 또한 더욱 확장될 수 있음을 보여주었다.

[그림 3-14]에 표시된 것처럼 NVIDIA의 Grace-Hopper GPU 기반 vDU, 기성품 4T4R RU, Keysight에서 제공하는 채널 에뮬레이터로 구성된 실시간 OTA(Over-The-Air) 테스트베드에서 시연되었으며, 제안된 실시간 AI 기반 채널 추정 솔루션의 안정성과 성능 개선을 확인하였다.



출처: AI RAN Alliance

[그림 3-14] AI 기반 PUSCH 채널 추정 테스트베드 구조

시연은 기지국(gNB)과 단말(UE) 사이에 현실과 유사한 무선채널 에뮬레이터를 사용하여 대상 커버리지 SNR 영역 내에서 수행되었으며, UL iperf 테스트는 전체 베퍼 트래픽 시나리오를 에뮬레이션 하는 데 사용되었다. 종래 비 AI 채널 추정 알고리즘이 먼저 테스트되었고 상향링크 MAC 트래픽 처리량은 10Mbps 이하를 보였다.



출처: AI RAN Alliance

[그림 3-15] PUSCH Throughput 결과

AI 채널 추정 알고리즘이 실험실(Lab) 학습모델기반으로 동일한 SNR 영역에서 수행되었을 때 상향링크 MAC 처리량이 ~15Mbps까지 증가하여 종래 비 AI 채널 추정 알고리즘에 비해 최대 50% 더 높은 처리량을 달성할 수 있음을 보여주었다. 또한, 온라인 학습모델을 기반으로 동일한 SNR 영역에서 수행되었을 경우 AI 채널 추정 알고리즘은 종래의 비 AI 채널 추정 알고리즘보다 2배 더 높은 최대 20Mbps의 상향링크 트래픽 처리량을 달성할 수 있었다.

3.4.3 모빌리티 인식 AI 기반 간섭 완화 및 에너지 절약을 위한 5G 빔포밍 기법

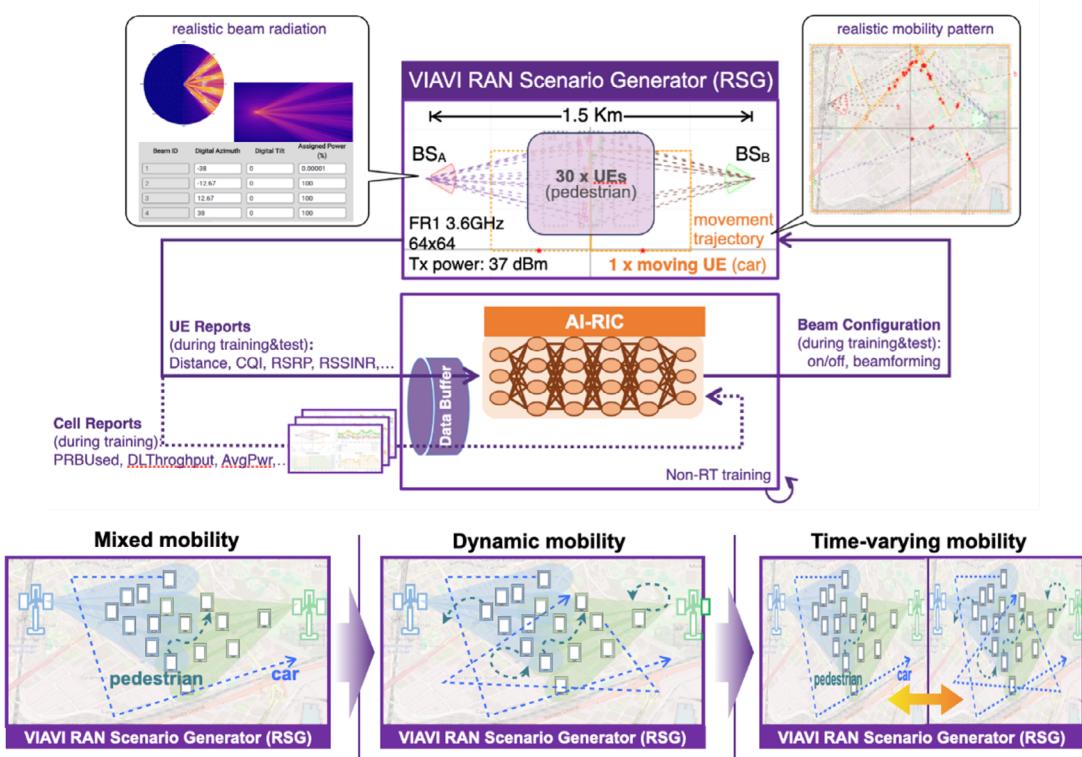
본 연구는 사용자 단말의 모빌리티 패턴을 학습하고 빔포밍을 최적화하여 RAN 다운링크(DL) 간섭 완화 및 에너지 절약을 위한 AI 응용프로그램을 개발하는 데 중점을 두었다. 사용자 단말의 이동성으로 인해 이전 CSI를 현재 시점에서 사용하기에 부정확하고, 이를 그대로 사용하는 경우 여러 기지국이 간섭 또는 중복 전송을 자주 겪을 수 있다. 따라서 완벽한 CSI 정보는 아니더라도 AI 기반 학습과 테스트 모두에 현실적인 RAN 시나리오 데이터를 적용하여 빔포밍과 전력 할당을 동시에 최적화하여 간섭 완화 및 에너지 절약 효과를 얻을 수 있는 모빌리티 인식 AI 기반 RAN 지능형 컨트롤러(AI-RAN Intelligent Controller AI-RIC)를 개발하였다.

[그림 3-16]과 같은 VIAVI RAN 시나리오 생성기(RAN Scenario Generator, RSG)를 사용하여



생성되는 현실과 유사한 모빌리티 및 빔 방사 데이터를 활용하여 지능형 컨트롤러(AI-RIC)을 학습하고 테스트하였다. 구체적으로, 64x64 빔포밍 어레이를 사용하여 FR1(3.6GHz)에서 작동하는 2개 기지국에 massive MIMO 시스템을 적용하였다. 본 연구에서는 다음과 같은 3가지 서로 다른 모빌리티 시나리오를 고려하였다.

- 혼합 모빌리티: 1.5km 범위에 분산된 하나의 빠르게 움직이는 단말(예: 자동차)과 30개의 정적 또는 느리게 움직이는 보행자 단말의 조합.
- 동적 모빌리티: 고정된 분포에서 샘플링된 무작위 모빌리티 패턴을 갖는 모든 단말.
- 시간 가변 모빌리티: 모든 단말이 주어진 분포에서 샘플링된 무작위 모빌리티 패턴을 갖는 정기적인 간격의 시간 가변 모빌리티 분포.



출처: AI RAN Alliance

[그림 3-16] RSG 기반 AI-RAN training 및 UE mobility patterns

지능형 컨트롤러(AI-RIC)는 가벼운 신경망(Neural Network, NN)으로 구성되며 빔포밍 및 빔 온/오프 매개변수와 같은 동작을 수행하는 동안 강화 학습(Reinforcement Learning, RL)을 사용하여 학습한다.



3. AI-RAN 서비스 및 기술 동향

지능형 컨트롤러(AI-RIC)를 학습시키기 위해 VIAVI RAN 시나리오 생성기(RSG)는 신경망의 입력 데이터를 생성하고, NN의 출력 동작에 대한 보상을 결정하고, 이러한 동작을 다음 신경망의 입력 데이터에 대한 변경 사항에 반영하여 강화학습의 환경을 구성한다. 학습하는 동안 RSG는 신경망의 입력으로 UE 보고 데이터(예: 거리, CQI, RSRP, RSSINR)를 제공하고 신경망 동작에 대한 보상으로 셀 보고 데이터(예: PRB 사용, 다운링크 처리량, 평균 전력 소비량)를 제공한다. 이러한 단말 보고 및 셀 보고 데이터 세트는 비실시간 배치 학습에 사용된다. 학습이 완료된 후 테스트하는 동안 온라인 단말 보고 데이터가 학습된 신경망에 입력되고 RSG에서 생성한 셀 보고 데이터를 기반으로 성능이 평가된다.

RSG는 학습과 테스트 전반에 걸쳐 단말 보고서와 셀 보고서 데이터가 생성되는 이전에 언급된 3가지 서로 다른 이동성 패턴을 에뮬레이션한다. 학습과 테스트의 계산적 복잡성을 줄이기 위해, 상위 레벨 동작이 빔 켜기/끄기를 결정한 다음, 빔포밍 매개변수를 최적화하는 하위 레벨 동작이 이어지는 계층적 RL 프레임워크를 사용한다. 이 순차적 작업은 후자의 동작 결정 공간을 줄여 복잡성 감소와 저지연 구현을 가능하게 한다.

또한 상위 레벨 동작은 이동성에 따라 비실시간일 수도 있어서, 하위 레벨 동작만 준실시간이 되도록 운영 시간 척도(operational time scales)를 분할할 수도 있다. 결과적으로 지능형 컨트롤러(AI-RIC)는 일괄적인 강화학습의 학습과 상위 레벨 빔 켜기/끄기 결정을 위해 사용되는 2개의 비실시간 rApp과 빔포밍 결정을 위한 1개의 준실시간 xApp으로 구성된다.

본 연구에서는 언급된 3가지 서로 다른 이동성 패턴에 따라 계산(computation) 성능측면에서 최대 20% 향상되었고, 통신 에너지 절약(Communication Energy Saving)측면에서는 평균 하향링크 트래픽 처리량이 최대 30% 증가함을 보여 주었다.