

신용카드리뷰

The Credit Card Review

Vol 19-1(2025년 03월)

LLM의 시장 영향: 시나리오 플래닝을 통한 AI시장 구조 전망

최 윤 정

SK경영경제연구소 수석연구원 & 한성대학교 경제학과 박사과정

김 시 언

주식회사 위너아이 이사

여 효 성

한성대학교 경제학과 조교수



한국신용카드학회
Korean Creditcard Academic Society

www.kci.go.kr

LLM의 시장 영향: 시나리오 플래닝을 통한 AI시장 구조 전망*

최 윤 정**

SK경영경제연구소 수석연구원 & 한성대학교 경제학과 박사과정

김 시 언***

주식회사 위너아이 이사

여 효 성****

한성대학교 경제학과 조교수

〈Abstract〉

본 연구는 LLM 시장 참여자간의 주도권 경쟁이 미래 AI 시장 구조에 미치는 영향을 시나리오 플래닝을 통해 탐색하였다. LLM은 방대한 데이터 학습을 기반으로 인간의 직관에 비견되는 속도로 높은 수준의 결과물을 내어놓았다. 이러한 LLM의 높은 잠재력과 시장 기대로 인해, 일부 부작용이 노출되었음에도 불구하고 이를 보완 기술이 등장하며 빠르게 AI 생태계를 형성해가고 있음을 목도하고 있다. 하지만 LLM 모델 자체 개발에 드는 막대한 비용과 고급 GPU의 공급 한계로 인하여 모델 개발사는 자본력을 가진 일부 기술 선도 회사에 한정되는 특징을 가진다. 이러한 기술적 특수성과 주요 사업자가 보유한 자원, Infra 조건, 환경적 요인에 기반한 시나리오 분석을 통해 미래의 LLM을 둘러싼 경쟁이 AI시장의 경쟁 구도에 어떤 영향을 미칠 것인가에 대한 예측가능한 미래상을 구성해보았다.

시나리오에 영향을 미치는 주요 변수를 식별하고, 사업자별 기술개발 방향성과 경쟁에 영향을 미치는 Infra 및 환경 요소를 확인하기 위해서 국내외 전문가와 인터뷰를 진행하였다. 현재 가장 많은 전문가와 투자자들이 지지하는 Base Scenario가 영향력이 높은 7개의 변수로 분화될 경우 도달 가능한 최종 시나리오들에서 AI 시대에 필요한 정책 측면의 기회와 위협을 논의한다.

핵심 단어: AI, LLM, 시나리오 플래닝, 미래 시장 전망

* 본 연구는 한성대학교 교내학술연구비 지원과제임.

** (제1저자) SK경영경제연구소 수석연구원, 서울시 종로구 수송동 율곡로2길 19 수송스퀘어 14F, E-mail: yj.choi@sk.com

*** (공동저자) 경기도 고양시 성신로 97 상가동 104호, Tel: 02)3275-3780, E-mail: kseon0704@naver.com

**** (교신저자) 서울 성북구 삼선교로 16길 116 한성대학교 521호, Tel: 02)760-4064, E-mail: hsyee@hansung.ac.kr

I. 서론

LLM(Large Language Model, 대규모 언어 모델)의 등장과 함께 AI 기술은 경제, 사회, 문화, 교육 등 다양한 분야에 급격한 변화를 일으킬 것으로 기대되고 있다. LLM은 기존의 특수 목적에 맞춰 최적화된 알고리즘과 달리, 범용성을 갖춘 AI 모델로 대중에게 그 능력을 보여주었다. 특히, 막대한 데이터를 기반으로 가장 관련성 높은 답변을 도출하는 능력은 인터넷 검색, 정보 요약 등의 영역에서 AI의 활용 가능성을 크게 확장시켰고, 이를 통해 일반 대중들도 AI 기술에 대한 기대와 관심을 높이고 있다.

또한, LLM의 근간이 되는 Transformer 기반 기술이 언어를 넘어서 이미지, 영상까지 적용된 멀티모달 기술이 연달아 등장하면서, AI가 디지털 너머 세상뿐 아니라 물리적 세상의 데이터를 수집하고 이해하는 데까지 진전되고 있다. 이는 기존의 생활, 일, 관계 등에서 우리가 당연하게 여기고 있는 일상적인 규칙이 무너질 것이라는 불안감과 동시에, 새로운 혁신적 변화를 가져올 것이라는 기대를 불러일으키고 있다.

그러나 LLM의 잠재력을 디스토피아적인 미래나 또다른 산업 혁명으로 연결하기까지는 아직 많은 변수들의 조합이 필요하다. LLM 모델의 성능은 높은 잠재력에 반해 그 정확도는 당장 실행 가능한 서비스나 프로세스에 적용하기에는 불안감이 높다. 게다가 LLM의 계산에는 다량의 초고성능의 칩셋, 에너지 소비량 100MW 이상의 하이퍼스케일 데이터 센터 등 높은 수준의 물리적 자원을 필요로 하기에 아직 사업자들은 LLM의 사업성을 담보하기 위한 비용 최적화에 어려움을 겪고 있다.

대부분의 논문에서 LLM으로 인한 경제 및 사회 측면의 변화를 AI가 시장에서 신뢰도를 얻어서 일과 생활을 변화시키는 결과론적 모습을 상정하고 있다. 하지만, 그 과정에서 주요 변화 동인이 무엇인지, 그 동인이 어떤 변화를 추동하고 시장에 어떤 결과를 가져오는가에 대한 분석은 부족하다. 이에 따라 AI 모델을 중심으로 시장이 소수 권력에 집중될 것인지, 다양한 사업자가 경쟁할 수 있는 시장으로 진화할 것인지에 대한 논의도 충분하지 않다. 따라서, 본 연구는 본 연구는 주요 시장 요소들을 식별하고, 그들이 시장 경쟁과 권력 재편에 미치는 영향을 분석하여 4개의 시나리오를 도출하였다.

II. 선행 연구 및 방법론

LLM은 일반 대중들에게 AI 기술의 유용성을 전파한 첫 번째 AI 모델으로 평가할 수

있다. 수많은 데이터에 기반한 검색, 요약, 질의응답 등의 기능은 AI가 대중적으로 서비스화할 수 있는 가능성을 제시하였으며, 이로 인해 높은 개발 비용에도 불구하고 높은 투자를 이끌어 내고 있다. 짧은 시간 내에 선도 기술 사업자들이 앞다투어 모델을 거대화하며 향상된 모델을 내놓고 있으며, Meta와 같은 기업은 거대 모델 경쟁과는 차별화되는 오픈 소스의 경로를 제시하기도 했다. Humza (2023)은 빠르게 업데이트되는 모델 중 40여개의 모델의 구조, 학습 및 추론 방식, 평가 방식, 응용 서비스 등으로 분석하여 정리하였다. 특히, LLM은 그 거대한 규모 때문에 구조적 변화나 학습 전략의 작은 변화만으로도 성능이나 안정성에 큰 영향을 가져올 수 있음에 주목한다. 각 개발사들은 MOE(Mixture of Experts), Parallel Attention, Positioning Encoding 등을 이용한 구조적 변화를 시도 중이며, 데이터 학습에 있어서 학습 비용을 최소화하면서 보다 나은 성능을 도출하기 위한 Controllable Text Generation, Mode Switching 등의 전략을 사용하고 있다. 해당 연구에서 현재까지의 LLM의 성능을 좌우하는 요소와 개발사 입장에서 제약 사항으로 작용하는 요소를 구체적으로 비교해볼 수 있다. 특히, Mckinsey (2023)의 보고서를 통해 OpenAI, Google, Meta의 모델 사업자와 독점적 칩 공급사 NVIDIA의 시장 선점 효과를 확인할 수 있다.

LLM 시장 확대에 가장 큰 걸림돌이 되는 것은 그 성능의 문제라고 할 수 있다. 최근까지도 Gogulamudi (2024) 등의 연구에서 연관성 있는 출력값을 선택하는 알고리즘의 태생적 구조로 인한 오류 출력의 문제는 개선이 쉽지 않음을 지적하고 있으며, 이는 서비스화의 걸림돌이 되고 있어 기술적 돌파구가 필요한 상황이다. 이를 보완하기 위한 기술이 모델의 후방에서 생태계를 형성하고 있으며, 대표적으로 입출력에 해당하는 정답을 넣은 데이터베이스를 참고로 활용하는 RAG(Retrieval Augmented Generation, 검색 증강 생성)가 주목받고 있다(Lewis, P. et al, 2020). 이는 정답 데이터베이스를 구성하는데 법, 의료 등의 전문 영역에서 모델의 오류를 제어할 수 있는 설계를 제공한다. 이렇게 LLM에 RAG를 통한 논리적 근거를 추가하는 설계는 Kahneman (2011)에서 설명한 인간의 인지 시스템을 구성하는 시스템 1과 시스템 2의 조합으로 이해할 수도 있다. 막대한 데이터에서 연관성에 기반해 빠르게 통계적 대답을 내는 LLM은 직관적 답을 내는 시스템 1을 모방하고 있으며, 정확한 질의응답으로 구성하여 논리성을 지원하는 RAG는 시스템 2에 가깝다. AI가 인지 구조를 모방하며 기능을 개선해 나가는 모습을 우리는 AI 시장의 생태계 확장에서 목도하고 있는 것이다.

또 다른 제약 사항 중 하나는 높은 계산 비용이다. LLM 알고리즘은 막대한 데이터 속에서 통계적 맥락을 찾아내고 가장 연관성이 높은 조합을 만들어 내는 것이기 때문에 입력된 토큰마다 막대한 양의 연관성 분석을 위한 계산이 요구된다. 또한 Hoffmann (2022)의 연구에서는 LLM 성능 최적화는 단지 파라미터의 수만 늘리는 것이 아니라 파라미터 수에 적합한 수준으로 데이터양을 늘려야 효과를 얻을 수 있다는 것을 확인하였

고, 이후 LLM의 품질 확보에는 막대한 데이터와 이를 가능케 하는 GPU, 전력 수급 등의 데이터센터 인프라가 핵심 요소로 등장하였다.

전력 문제 개선을 위해 모델의 알고리즘 자체를 효율화하는 연구도 활발하다. 특히, Gu and Dao (2024)은 모든 토큰을 계산하지 않고 맥락을 별도로 제공하여 토큰을 압축하는 효과를 내는 State Space Model(SSM)을 이용한 Mamba 모델로 주목을 받았다. AI21 Studio라는 개발사는 Mamba를 기존 LLM과 혼합(Mixture of Experts)하여 사용하는 Jamba를 공개하였으며(Lieber, O. et al, 2024), 현재 LLM 모델 강자인 빅테크도 내부적으로 테스트 중인 것으로 알려져 있다.

거대 모델 사용에 따른 서비스 사업자 부담을 해소하기 위해 오픈소스를 활용하는 연구도 활발하다. Schillaci (2024)는 NVIDIA가 제공하고 있는 거대 모델용 서버칩 H100을 사용하는 거대 모델을 서버로 이용하는 것이 아닌 오픈 소스를 현장(온사이트)에 설치하는 방식이 사이버 보안, 교육, 비용 측면에서 유리함을 주장하였다. 하지만 오픈 소스를 직접 설치해서 서비스화하는 시도는 높은 초기 투자 비용, 전문 지식 필요, 성능 최적화에 따른 성능 차이 등의 문제로 개별 서비스 업자가 실행하기에는 아직 한계가 있다. 이를 해결하기 위해서는 전후방에서 이를 효율적으로 지원하는 생태계가 형성될 필요가 있다.

LLM은 향후 AI 기반 서비스의 핵심 기술로 자리 잡을 것이며, 이에 따라 LLM 개발 경쟁은 미래 경쟁 구도를 결정짓는 중요한 요인으로 작용할 것이다. 시장에서는 이미 LLM의 기술적 한계와 제약 사항을 어느 정도 인지하고 있으나, 급속한 기술 발전으로 인해 이 역시 선도 기술 기업들이 해결할 것이라는 막연한 기대도 존재한다. 그러나 이러한 기대는 불확실성이 크며, 특히 LLM의 제약 사항들이 AI 시장의 경쟁 구도에 미칠 영향을 종합적으로 분석한 연구는 부족하다. 이에 본 논문은 LLM 기술의 성능, 비용, 서비스 시장의 핵심 동인을 바탕으로 미래 시장을 전망하고, 이를 예측하기 위해 시나리오 플래닝 기법을 활용하여 시장 전개 방향을 제시하고자 한다. 이는 미래 전망에 있어 핵심 동인과 변수를 식별해서 가능한 전개 방향을 전망하고 이에 기반하여 주체별 기회와 위협을 분석할 수 있는 기법이다. 시나리오 플래닝 기법은 미래의 불확실성이 큰 상황에서 영향력이 큰 변수들에 기반하여 미래 변화 방향의 분기를 만들고, 의사결정자에게 미래에 대한 마인드맵을 제공함으로써 불확실한 경계를 식별하도록 하여 전략적 판단에 기여한다(Lindgren, 2003). Korinek (2023)은 AGI의 경제적 영향을 세 가지 시나리오로 분석하며, 5~10년 내 AGI가 도달할 가능성에 대해서 1) 오히려 일자리가 창출, 2) 20년내에 AGI가 도달하며 노동가치가 점차적 하락, 3) 2)와 유사하지만 5년내 AGI 도달후 빠르게 노동시장에 영향을 미치는 3가지의 시나리오를 분석하였다. 본 연구는 노동시장의 변화가 아닌 사업자가 경쟁 지혜의 변화를 다양한 시나리오로 탐구한다.

시나리오 플래닝은 통한 미래상 개발은 다음의 다섯 단계를 거치게 된다. 1단계는 문

제를 식별하는 단계로 무엇을 다루고 무엇을 밝히고자 하는지 경계를 정하는 것이다. 본 연구는 LLM 모델을 둘러싼 경쟁을 통해 AI시장의 미래를 전망하는 것을 목표로 한다. 두 번째 단계는 요인 파악 및 분석 단계로, 의사 결정에 영향을 줄 수 있는 미래 요소를 수집하고, 세 번째의 선정단계에서 미래의 분기를 만드는 주요 동인과 이에 영향을 주는 변수들을 선정한다. 네 번째 단계에서는 이러한 동인과 변수를 조합하여 시나리오를 전개하는데 불필요한 분기들을 정리하여 최종 시나리오를 정하게 된다. 마지막으로 최종 시나리오를 통해 전략적 시사점을 도출하게 된다(울프 필칸, 2009).

III. 주요 변수 식별

현재의 AI 시장은 강력한 모델과 칩셋이라는 두 축을 중심으로 견인되고 있으며, 본격적 시장 확대의 키인 서비스 시장은 아직 개화되기 전이다. AI를 위한 계산 인프라에 필수 요소인 고성능 칩셋은 NVIDIA가 90%의 압도적인 경쟁 지위를 보유하고 있으며, 이 독점적 지위는 향후 5년 이상 지속될 것이라는 게 전문가들의 분석이다. 모델과 칩셋 경쟁에 진입하기에는 기술과 자본의 진입 장벽이 높으며 강력한 선도 모델 사업자가 칩셋을 확보하는 시장 구조 때문에, 서비스의 개화 또한 결국 모델 사업자에게 종속되지 않겠냐는 예측이 현재의 베이스 시나리오를 구성한다.

베이스 시나리오는 현재의 기술 고도화 방향성과 주요 사업자의 차별적 지위가 지속된다면 전개 가능하며 시나리오 1로 연결된다. 베이스 시나리오의 분기점으로 찾기 위해서는 현재의 경로에서 분기 가능한 주요 동인을 식별해야 하고자 한다. 이 분기점이 별 영향력이 없다면 시나리오 1으로 수렴한다고 가정한다.

첫번째 핵심 동인은 모델의 지속 고도화 여부이다. 현재의 LLM 모델은 태생적으로 오류 출력 가능성이 있는 알고리즘이지만 OpenAI와 같은 사업자는 막대한 데이터와 모델을 여러 개 배치하는 구조적 변화 등을 통해서 이를 돌파하고자 한다. 아직 명확한 논리 시스템(Reasoning)에 대한 기술적 돌파구는 명확하지 않지만 이를 해결하는 모델 혁신이 있을 것인가 역시 첫번째 분기의 주요 판단 요소가 될 것이다. 이러한 모델의 지속적 고도화의 필수 요소로는 인프라 자원 측면에서 모델의 진화를 뒷받침하는 칩셋의 진화가 지속될 것인지, 인터넷 상에서 수집할 수 있는 데이터는 고갈이 노정된 가운데 전문 데이터 및 합성 데이터를 통해 모델 성능 개선에 필요한 양질의 데이터가 충분히 공급될 것인가도 포함된다.

두번째 핵심 동인은 모델 운영의 경제성 확보 여부이다. Huawei (2023)은 AI 연산량 증가 폭은 0.5년에 2배 수준으로 증가하지만 GPU 전력 효율 개선은 3.5년에 2배 수준 개선에 불과하여 현재 연산 비용이 급격히 증가가 불가피함을 분석하였다. 전력 문제 개

선을 위해 다양한 개선안이 나오는 중으로 State Space Model(SSM) 등의 새로운 구조의 LLM이 시도되고 있으며, 일부 서비스 사업자는 공개된 오픈 소스를 특정 목적 용으로 학습시킨 작은 모델(SLM: Small Language Model)을 이용하는 서비스로 비용의 문제를 우회하려고 시도 중이다. 이러한 흐름에서 OpenAI, Google 과는 달리 오픈 소스로 접근한 Meta가 시장이 다양하게 분기하는 계기를 제공하고 있다. 또한 오픈 소스는 모델을 서비스 맞춤형으로 최적화하기 위한 운영 최적화 및 효율화를 겨냥하는 후방 시장을 보다 두텁게 만들면서 경제성을 확보하는 데에도 기여하고 있다.

실제 서비스가 활성화되는 단계의 모델 사용 비용을 고려하면 디바이스 구매자가 AI 계산 비용을 분담하는 구조도 예상 가능하다. AI 모델의 비용구조는 과거 디지털 서비스와는 달리 한계 비용 제로의 구조가 불가능하기에 무료로 우선 소비자를 모으고 나중에 유료화하는 기존 플랫폼 전략을 펼치기 쉽지 않다. 심지어 기존 수익 구조가 탄탄한 빅테크에게조차 이는 고민의 대상이다. 따라서 지불의사가 확고한 서비스 출현이 지연된다면 디바이스의 칩셋으로 구동되는 초소형 모델을 중심으로 사용하는 온디바이스 AI가 주가 되는 전망도 가능하다. 이는 모델 비용을 디바이스 구매 비용에 포함시키는 구조가 된다.

세번째 핵심 동인은 서비스 혁신을 통한 시장화의 여부이다. 아직은 AI 모델의 범용적 활용성을 검증하는 서비스의 극초기 단계로 검색, 요약, 챗봇, 번역 등의 영역에서 소구 지점을 발견하고 있다. 이런 범용 서비스가 소비자들의 지불 의사를 이끌어 내지 못할 경우, 특정 분야의 특화된 서비스를 중심으로 수익이 발생하면서 서비스향으로 최적된 SLM(Small Language Model)이 다양하게 존재하는 시장으로 진화할 수 있다. LLM을 이용한 다양한 개발 프레임워크를 제공하며 주목받은 Hugging Face의 CEO는 '더 작고, 빠르고, 맞춤화된 모델이 실제 사용 사례의 99%일 것이다. 보통 백만불짜리 레이싱카가 필요없기 때문이다'라고 소형 전문화된 모델의 서비스 시장 확대 가능성을 긍정적으로 평가한 바 있다(Delangue, 2024). 하지만 자본 집약적이고 빠른 기술 혁신으로 AGI(Artificial General Agent)의 독점적 지위를 예견하는 이들은 현재의 주류 의견이다. 거대 단일 모델이 다양한 서비스에서 월등한 성능을 내면서 진화하는 AGI는 컴퓨터와 모바일폰의 OS 처럼 새로운 플랫폼으로 진화하여 모든 3rd Party 사업자와 소비자를 흡수하는 미래를 그리고 있다. 그리고 이러한 전망은 정책적으로 AI 역량을 국가적으로 육성하고 보호하려는 Sovereign AI 추세가 등장하는 배경이기도 하다.

현재의 고객 소구가 명확함에도 불구하고 높은 비용으로 인해 서비스 시장 개화가 지연될 경우 투자 시장이 위축될 가능성도 이 세번째 동인에 포함된다. WEF(2025) 연구에 따르면 OpenAI의 ChatGPT에 대한 단일 쿼리는 약 2.9Wh의 전기를 소비하며 이는 Google 전통적 키워드 검색의 약 10배에 해당한다고 밝혔다. 또한, Schoon (2023)는 알파벳의 회장인 John Hennessy와의 인터뷰를 통해 비용 측면에서 대규모 모델의 검색이 기

존 검색의 10배까지 증가할 수 있음을 언급하며 전체 검색을 AI로 대체하지 못하는 이유를 설명하였다. 강력한 B2B 사업자인 MS도 기존 MS365 패키지에 인별 추가 비용을 추가하는 구조로 Copilot이라는 AI 서비스를 내놓았지만 기업별 데이터 보안 구조부터 비용까지 서비스 확장에는 많은 허들을 마주하고 있다. 투자 자본에 의존적인 스타트업의 입장에서는 서비스에 대한 기대감이 떨어질수록 투자자들이 회의적인 기조로 돌아설 위험에 대한 우려가 있는 것도 고려해야 할 요소이다

<그림 1> 핵심 동인과 하부 변수



IV. 시나리오 전개

4.1 시나리오 1(AGI의 플랫폼화)

이 시나리오는 현재의 선도 기술 사업자의 기술 주도권이 유지될 경우로, 지속적인 모델의 성능 개선이 이루어지면서 독보적인 모델을 보유한 사업자가 LLM 자체를 플랫폼으로 서비스 시장의 주도권까지 가져가는 시나리오이다. 이 시나리오는 특정 사업자가 압도적인 LLM 성능을 기반으로 시장에서 독점적인 지위를 획득하는 시나리오로서 해당 사업자가 비용에 대한 제어력도 어느 정도 획득하게 됨에 따라 비용 개선 여부에서의 분기는 중요도가 떨어진다.

반도체 사업자의 기술 혁신이 지속되면서 무어의 법칙을 넘어설 정도의 성능을 보유한 GPU를 지속 공급하고, 데이터 가치와 거래 방식에 대한 사회적 합의가 이루어지며

데이터 유통 생태계가 정비된다고 상정한다. 또한 합성데이터 기술이 발전함에 따라 모델 학습에 필요한 데이터가 양적, 질적으로 충분하게 공급된다. LLM 기술을 주도하는 사업자는 이러한 인프라와 데이터의 지원 하에 모델의 서비스화 제약 조건으로 언급되던 오답 출력 현상을 높은 수준으로 개선하고 스케일업을 지속하면서 타 사업자와의 격차를 높여간다. 또한 LLM 모델을 플랫폼으로 서비스 사업자에게 제공할 수 있는 LLM 개발 환경을 독점적으로 제공하면서 LLM을 AI 서비스 생태계의 엔진으로 활용하여 다양한 서비스가 개발되는 생태계를 조성해 나간다. 이런 조건에서 독점적 기술적 입지를 확보한 사업자는 인터넷 생태계의 가치를 급격하게 빨아들이면서 강력한 AI 플랫폼 사업자로 자라잡는다.

4.2 시나리오 2(Big-Tech 주도)

이 시나리오는 반도체 성능이 한계에 도달함에 따라 LLM의 파라미터 거대화 경쟁도 한계를 맞이하게 된 이후의 전개이다. Computing 비용의 개선이 이루어지지 않아 신규 서비스가 개화하지 못하고 기존 서비스 강자가 기존 시장 지배력을 기반으로 AI 비용을 보조하며 AI 서비스는 기존 서비스의 부가 가치 제공 수단으로 주로 활용된다. 이에 따라 기존의 빅테크가 AI 시장에서 보다 강화된 입지를 갖게 된다.

반도체 성능의 진화가 정체되면서 모델 거대화는 물론, 시각 정보 등을 활용하는 멀티모달 모델도 성능 정체기에 접어든다. 즉, LLM 모델은 AIG급 성능까지는 도달하지 못한다. 모델의 한계가 노정되면서 시장은 모델 경쟁 보다는 빠르게 투자를 수익화하기 위한 서비스 경쟁으로 접어들지만 Computing 비용 자체는 개선되지 않으면서 스타트업의 신규 AI 서비스 창출이 제한된다. 반면 빅테크의 경우 기존 서비스를 통해 확보한 서비스 역량에 AI의 가치를 부가하는 방식으로 서비스를 혁신해가며 기존의 시장 지배력을 강화한다. 게다가 빅테크는 직접 디자인한 반도체로 서비스와 모델 비용간의 수직 통합에 의한 효율화를 진행하면서 비용 개선에도 유리한 입지로 AI 서비스 경쟁을 주도한다.

4.3 시나리오 3(Specialized AI 전성)

이 역시 모델의 성능 개선이 한계에 다다르면서 LLM이 AGI급 성능까지 도달하지 못하는 시나리오이다. 이 시나리오에서 이러한 성능 개선의 실패는 고객들의 검색, 요약, 일반 질의 등의 범용 서비스에 대한 지불 의사를 견인하지 못하는 원인이 된다. 다만, 시나리오2와는 달리 역량있는 신규 서비스 사업자들이 상대적으로 작은 모델인 SLM을 이용하면서 비용 최적화에 어느 정도 성과를 보인다. 점차 고객들은 요약이나 검색 같은 범용 서비스보다는 법률, 의료, 교육 등의 특화 서비스에 지불 의사를 보이면서 각 서비스 사업자가 SLM과 전문적 데이터를 가지고 차별적인 서비스를 다양하게 창출하는 시

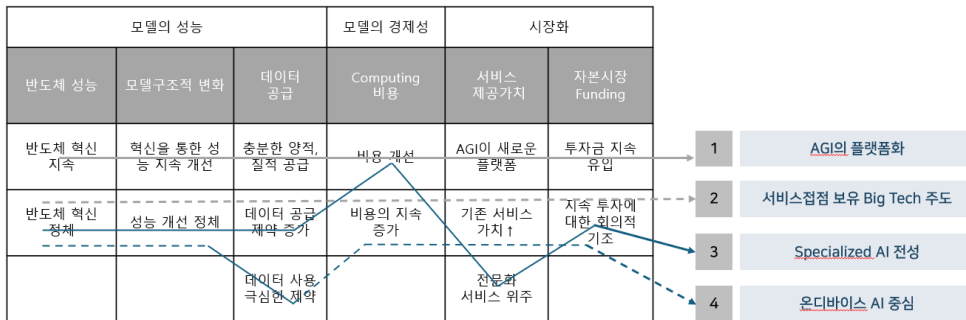
장이 조성된다. 가치 사슬에서는 거대 모델을 위한 고성능 반도체 공급자의 독점적 시장이 아닌 다양한 후발 사업자가 경쟁하는 중급의 반도체 시장의 활성화 되어 AI 시장의 Computing 비용 수준이 하락하게 된다. 이는 다시 전문 서비스 서비스의 비용 개선을 가져오는 선순환 구조를 강화한다.

4.4 시나리오 4(온디바이스 AI 중심)

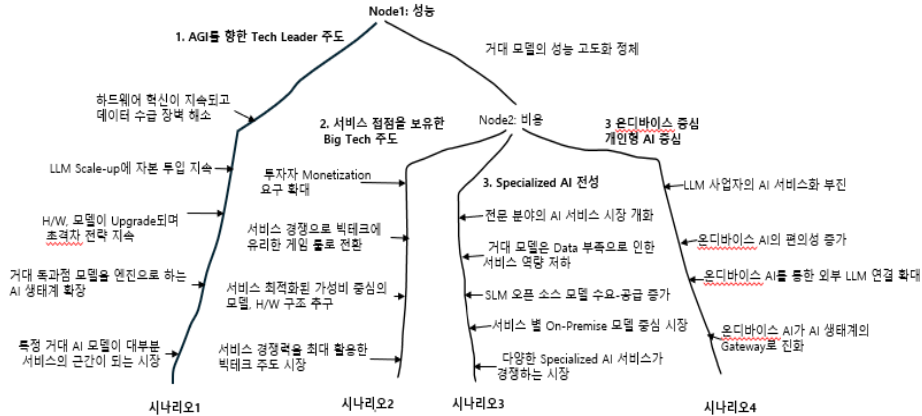
이 시나리오는 모델의 성능이 정체되면서 AGI에 대한 기대가 높지 않은 가운데, 개인 정보가 모델에 활용되는 것에 대한 사용자들의 거부감이 높아지거나 정책적으로 개인 정보 보호에 대한 강한 정책이 수반되면서 데이터 사용에 극심한 제약이 생기는 경우이다. 이러한 데이터의 제약은 결국 모델 성능의 저하로 연결되기 때문에 고객들의 서비스 이용 경험이 악화되는 악순환으로 이어진다. 게다가 Computing 비용은 개선되지 않는 상황이기 때문에 고객들이 돈을 내고 AI 서비스 사용할 의사는 더더욱 낮아지게 된다.

이런 와중에 애플 등의 모바일 OS업체는 모델이 디바이스 안에서 구동되어 데이터의 외부 유출 가능성이 없는 온디바이스 AI를 제공한다. 거대 모델인 LLM이나 서비스 업체의 SLM보다는 성능이 떨어지지만 일상적인 사용에서는 오히려 개인 맞춤형의 질의응답이 가능하기 때문에 고객의 소구 지점을 자극한다. 이를 적극적으로 활용할 수 있는 모바일 OS용 API를 공급하면서 서비스 사업자들도 해당 API를 이용한 AI 서비스를 출시하고 모바일 생태계가 AI 생태계로 확대되는 결과를 가져온다.

< 그림 2 > 핵심동인과 하부 변수에 따른 시나리오 분기



<그림 3> Scenario Delta Chart



V. 논의 방향

상기 4개의 시나리오는 배타적으로 전개되기 보다는 시기적으로 중첩되고, 일부는 전후의 인과 관계로 연결될 수도 있다. 예를 들어 OpenAI의 AGI 개발을 위한 막대한 투자와 구글의 고유 서비스와 AI 역량을 결합하려는 시나리오 1과 2이 공존하는 상황에서 막대한 투자와 전략에도 불구하고 3년 이내에 LLM의 성능적, 비용적 한계로 인해 대중에게 소구할 수 있는 서비스가 출시되지 않을 가능성도 있다. 이로 인해 투자 시장이 위축되는 가운데 일부 전문적인 서비스가 지불 의사를 이끌어내며 생존하면서(시나리오3), 현재의 IT 강자 구도가 AI를 통해 새로운 시장로 재편될 가능성도 있다. 해당 전문적인 서비스의 도달 시장이 넓지 않은 경우에는 지불 의사가 낮은 저사양의 일상 서비스가 대중화될 가능성이 있기 때문에 전문적인 서비스와 일상 보조 서비스로 양분화되는 시장을 예견할 수도 있을 것이다. 이는 시나리오 3과 4가 공존하는 시장이라고 볼 수 있겠다.

위의 4개 시나리오 중 1과 2는 후발 사업자 대비 압도적인 기술적 역량 또는 기존 서비스 역량을 바탕으로 하기 때문에 국내 AI 시장의 기술적 역량과 사업적 역량을 고려할 때 자생적인 AI 성장을 예상하기 쉽지 않다. 하지만, 시나리오는 결과론적인 미래 전망이 아니다. 우리는 전개 과정의 핵심 동인과 하부 변수들을 중심으로 국가 또는 사업자별로 도전해볼 수 있는 인사이트를 확인할 필요가 있다. 국내의 모델 역량 측면에서

절대적으로 불리한 시나리오 1과 2의 경우에도 그 과정에서 한국의 사업자가 접근할 수 있는 요인을 확인할 수 있다.

현재 국내 사업자들은 모델 개발 분야에서 두각을 나타내지 못하고 있다. 우선 자본 또는 기구축한 자산을 기본으로 하는 거대 모델 개발사의 기술적 해자를 돌파하기가 어렵다는 패배의식 또한 발견된다. 하지만 서비스 시장에서 실제로 AI의 잠재력을 경제적 가치로 실현하는 사업자는 단순히 모델 사업자로 귀결될 것인가를 다양한 관점에서 확인할 필요가 있다. 어떤 동인이 거대 사업자의 기술적 해자를 무너뜨리는 역할을 하는지 미래의 다양한 가능성을 탐색하기 위해, 어떤 신호를 포착했을 때 어느 방향으로 움직여야 할지 전략적 옵션을 고민해야 할 때 본 논문의 시나리오가 기여할 수 있을 것이다.

또한, 모든 시나리오에서 발견되는 비용 절감의 필요성은 인프라 경쟁에서 반도체 경쟁력, 에너지 기술 역량, 안정적인 통신망을 보유한 한국 시장에서 유리한 입지를 제공할 수 있다. 현재 AI시장을 둘러싼 인프라 경쟁의 양상은 AI데이터센터로 집중되고 있다. 특히, 주목할만한 것은 AI 데이터센터가 요구하는 전력 규모가 국가 전력 정책과 연계되어야만 가능한 GW 수준으로 확대되면서 각 국가 차원의 정책이슈로 등장하고 있다는 것이다. 이에 지역적으로 아시아 지역에서의 인프라 경쟁력을 바탕으로 타 영역의 경쟁력을 어떻게 견인할 수 있는가를 정책적으로 타진해볼 수도 있을 것이다.

이렇게 현재 모델 성능, 비용, 그리고 향후 서비스 시장 확대 가능성을 기준으로 분기에서 기회를 발견하는 시나리오 플래닝 접근은 국가와 기업의 AI 전략 차원에서 구체적인 실행 가능한 접근 방안을 마련하는 데 기여할 수 있을 것으로 생각된다.

참 고 문 헌

- 울프 필칸. (2009). **트렌드와 시나리오**. 서울: 리더스북.
- Delangue, C. (2024). Smaller, cheaper, faster, more customized models will cover 99% of use-cases. **Co-founder & CEO, Hugging Face**. LinkedIn. https://www.linkedin.com/posts/clementdelangue_ive-said-it-and-will-say-it-again-1-activity-7112524134395318273-T3z6?utm_source=share&utm_medium=member_desktop
- Gogulamudi, P. R., Kumar, Y. V. P., & Prakash, K. P. (2024). **Hallucinations in large language models (LLMs)** (In 2024 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)). <https://doi.org/10.1109/eStream61684.2024.10542617>
- Gu, A., & Dao, T. (2024). **Mamba: Linear-time sequence modeling with selective state spaces** (arXiv:2312.00752). arXiv. <https://arxiv.org/abs/2312.00752>
- Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ... & Gong, Z. (2022). **Training compute-optimal large language models** (Working Paper). DeepMind.
- Humza, N., Asad, U. K., Shi, Q., Muhammad, S., Saeed, A., Barnes, N., & Mian, A. (2023). **A comprehensive overview of large language models** (arXiv:2307.06435). arXiv. <https://arxiv.org/abs/2307.06435>
- Huawei. (2023). **Data center facility energy efficiency and infrastructure sustainability** (Huawei 2030 report). <https://www.huawei.com/en/giv/data-center-2030>
- Kahneman, D. (2011). **Thinking, fast and slow**. New York: Farrar, Straus and Giroux.
- Korinek, A. (2023). **Scenario planning for an AGI future**. IMF Finance & Development Magazine.
- Lewis, P., Perez, E., Piktus, A., ... & Kiela, D. (2020). **Retrieval-augmented generation for knowledge-intensive NLP tasks** (Proceedings of the 34th International Conference on Neural Information Processing Systems, 793, 9459-9474).
- Lieber, O., Lenz, B., & Bata, H., et al. (2024). **Jamba: A hybrid transformer-Mamba language model** (arXiv:2403.19887). arXiv. <https://arxiv.org/abs/2403.19887>
- Lindgren, M., & Bandhold, H. (2003). **Scenario planning: The link between future and strategy**. Springer.
- McKinsey. (2023). **The state of AI in 2023: Generative AI's breakout year**.
- Schoon, B. (2023). **Google's Bard AI could cost 10x as much as traditional keyword search**. 9to5Google.

The Market Impact of LLM: An Outlook on AI Market Structure Through Scenario Planning Approach

Yun Jung Choi*

Head Researcher, SK Research Institute, Korea & Ph.D. Student, Hansung University

Si Eon Kim**

Director, Winnereye Co., Ltd., Gyeonggi-do, Korea

Hyo Sung Yeo***

Assistant Professor, Department of Economics, Hansung University

—〈Abstract〉—

This paper explores the impact of competitive dynamics among LLM market participants on the future structure of the AI market through scenario planning. LLMs produce high-quality outputs at speeds comparable to human intuition based on extensive data learning. Despite some revealed side effects due to the high potential and market expectations of LLMs, we are witnessing the emergence of complementary technologies that rapidly shape the AI ecosystem. However, the substantial costs associated with developing LLM models and the limitations in the supply of advanced GPUs restrict model developers to a few capital-rich leading technology companies. By analyzing scenarios based on these technical peculiarities, the resources held by key players, infrastructure conditions, and environmental factors, we have constructed a predictable future regarding how competition surrounding LLMs will affect the competitive landscape of the AI market.

To identify key variables influencing the scenarios and to understand the technology development directions and infrastructure and environmental factors affecting competition, interviews with domestic and international experts were conducted. The Base Scenario, which is supported by the majority of experts and investors, discusses opportunities and threats in policy aspects from potential final scenarios that can be reached when it is differentiated by the seven influential variables.

Keywords: AI, LLM, Scenario planning, Future market outlook

<최초 투고일: 2025년 2월 14일>, <수정일: 2025년 3월 11일>

<게재 확정일: 2025년 3월 28일>

* Address: 14F, 19, Yulgok-ro 2-gil, Jongno-gu, Seoul, Republic of Korea, E-mail: yj.choi@sk.com

** Address: Sanggdong 104, 97, Seongsin-ro, Deogyang-gu, Goyang-si, Gyeonggi-do 10483, Korea, E-mail: dcd88@naver.com, Tel: +82-2-3275-3780

*** Address: 116, Samseongyo-ro 16-gil, Seongbuk-gu, Seoul 03016, Korea, E-mail: hsy eo@hansung.ac.kr, Tel: +82-2-760-4064