

자연어 처리



데이터분석본부 호남지원 선임연구원 **이은지** Tel: 062-951-7704 e-mail: eunji_lee@kisti.re.kr

KEY FINDING

1. 자연어 처리는 인간이 일상 생활에서 사용하는 언어인 자연어를 컴퓨터가 해석, 조작, 이해할 수 있도록 처리하는 기술이다.
2. 자연어 처리의 세계 시장 규모는 2022년 157억4,900만 달러이며, 연평균 25.7 % 수준으로 성장해 2027년에는 494억1,300만 달러가 될 것으로 전망된다.
3. 최근 AI 산업은 방대한 데이터와 파라미터를 활용하여 학습해 인간처럼 생각하고 추론이 가능한 초거대 AI 시대로 진입하였으며, 자연어 처리 분야에서는 대규모 언어 모델 학습으로 정확도가 견고해지면서 다양한 산업별 응용 서비스 확장이 기대된다.
4. 컴퓨팅 파워 및 서비스를 위한 인프라를 제공하거나, 대규모 언어 모델 등 AI 모델 개발의 경우 빅테크 기업이 주도해 나가고 있으며, 빅테크 기업의 AI 모델이 공개되면 중소 규모의 사업자는 구체적인 서비스와 애플리케이션을 개발하는 형태로 밸류 체인이 형성되고 있다.
5. 자연어 처리는 기반 기술, 플랫폼, 응용 서비스 등 융복합형 산업을 형성하고 있으며, 텍스트, 음악, 영상 이미지 등 생성형 AI에서 활용되어, 패션, 디지털 마케팅, 프로그래밍 등 초개인화된 콘텐츠 서비스 산업의 변화 동인으로 작용할 것으로 전망된다.

1) 시장의 개요

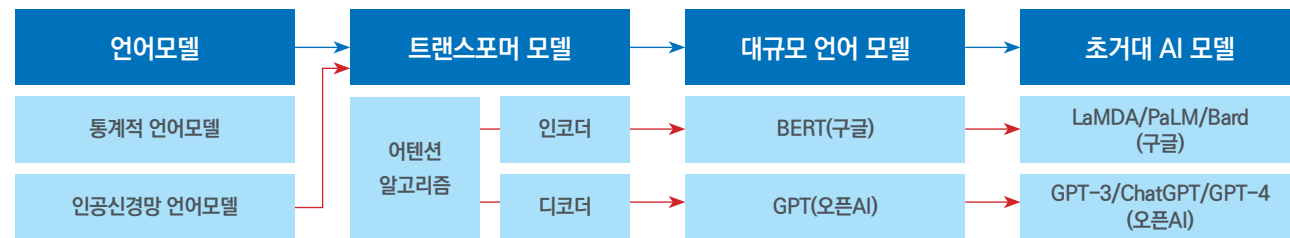
자연어 처리는 인간이 일상 생활에서 사용하는 언어인 자연어를 컴퓨터가 해석, 조작, 이해할 수 있도록 처리하는 기술을 일컫는다. 초기의 자연어 처리 기술로는 언어학적 지식에 기반해 분석하고 처리하는 규칙 기반 방식, 그리고 말뭉치(corpus)나 학습 데이터 내 언어적 특징(feature)의 출현빈도 등 특징정보를 활용하여 다양한 기계 학습 기법을 적용하는 확률 통계 방식이 이용되었다. 언어가 지닌 의미적 모호성,

중의성, 어순의 유연성 등과 같이 복합적인 문제 때문에 분석 성능에 한계를 가졌던 자연어 처리 기술은 딥러닝 기술의 발전과 적용으로 비약적으로 성능이 발전되는 계기를 마련했다. 딥러닝 기술은 특정 작업을 수행하기 위해 데이터에서 어떤 특징에 주목해야 하는지를 정의하는 기존 방식과는 달리 데이터로부터 다양한 특징들을 학습하여 스스로 주어진 문제를 해결하도록 학습을 추구하기 때문에 자연어 처리, 영상 처리, 음성 인식 등과 관련된 인공 지능(AI) 기술의 성능 향상에 기여하였다.

최근 자연어 처리 분야에서는 구글에서 발표한 트랜스포머 모델¹⁾을 기반으로 하는 대규모 언어 모델²⁾이 등장하면서 문장 분류, 질의 응답뿐만 아니라 복잡한 문장의 생성, 자연스러운 대화 등 대규모 언어 모델에 기반한 차별화된 AI 서비스가 공개되었다. 이러한 대규모 언어 모델은 사

전 학습된 언어 모델로서 처리하고자 하는 각종 작업에 맞게 미세 조정(fine-tuning)하는 전이 학습(transfer learning)을 통해 재학습해 사용하는 방식에서 점차 미세 조정을 거의 하지 않거나(few-shot learning), 전혀 하지 않는(zero-shot learning) 형태로 발전해 나가고 있다.

그림 1 언어 모델의 변화

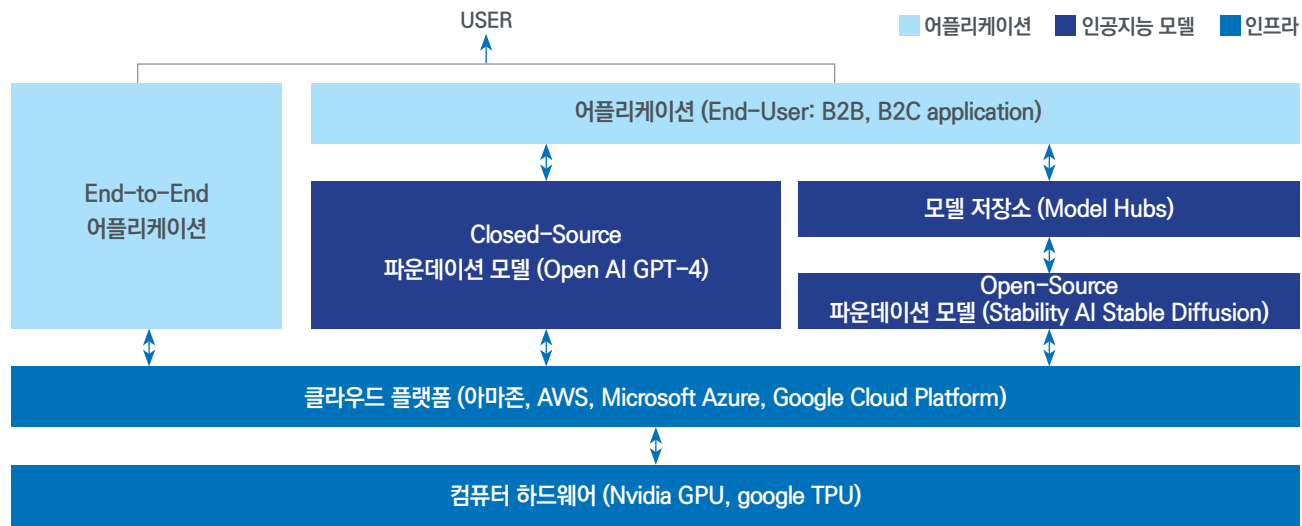


자료: 초거대언어모델의 부상과 주요이슈(2023), KISTI 재구성

대규모 언어 모델을 기반으로 하는 AI 산업의 밸류 체인은 '인프라', 'AI 모델', '애플리케이션'으로 대별된다. 컴퓨팅 파워 및 서비스를 위한 인프라를 제공하거나, 대규모 언어 모델 등 AI 모델 개발의 경우 빅테크 기업이 주도해 나가고 있으며, 빅테크 기업의 모델이 공개되면 중소 규모의 사업자는 구체적인 서비스와 애플리케이션을 개발하는 형태로 밸류 체인이 형성되고 있다.

자연어 처리 기술은 언어 모델 구축을 통해 타산업의 제품 및 서비스에 적용해 새로운 비즈니스를 창출할 수 있는 기반 기술로써 질의 응답, 문서 분류 및 요약, 문서 생성과 같은 지능형 정보서비스와 챗봇, 음성인식과 같은 대화형 인터페이스를 기반으로 하는 AI 서비스 개발에서 활용되며, 향후 디지털 휴먼 등 다양한 매체와 디바이스로 확장이 기대되는 분야이다.

그림 2 대규모 언어 모델 AI 산업 밸류 체인



자료: who owns the generative AI platform, a16z Enterprise(2023)

1) 트랜스포머 모델(transformer model) : 문장 속 단어와 같은 순차 데이터 내의 관계를 추적해 맥락과 의미를 학습하는 신경망
2) 대규모 언어 모델(Large Language Model) : 대용량의 말뭉치 데이터로부터 많은 파라미터를 이용해 언어의 문법과 의미를 학습해 다양한 응용 태스크에 범용적으로 적용할 수 있는 언어 모델

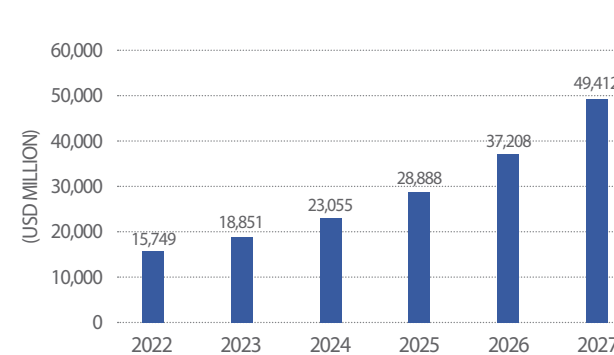
2) 시장 동향

| 시장 규모 및 전망

다양한 산업 분야에서 AI, IoT, AR 등의 첨단 기술을 서비스에 접목 하려는 시도가 계속되는 가운데 자연어 처리 기술은 사용자의 편의와

생산성을 높이는 다양한 애플리케이션 개발에 활용되고 있다. 자연어 처리 세계 시장 규모는 2022년 157억4,900만 달러이며, 연평균 25.7%로 성장해 2027년 494억1,300만 달러가 될 것으로 전망된다. 자연어 처리의 국내 시장 규모는 2022년 1억7400만 달러이며, 연평균 29.5%로 성장해 2027년 6억3,300만 달러가 될 것으로 전망된다.

그림 3 자연어 처리의 세계 시장 규모 및 전망(2022~2027)



출처 : Markets&Markets, Natural Language Processing Market, 2022

그림 4 자연어 처리의 지역별 시장 규모 및 전망(2022~2027)

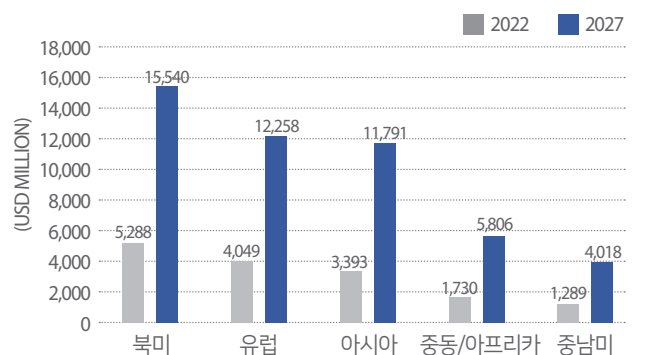
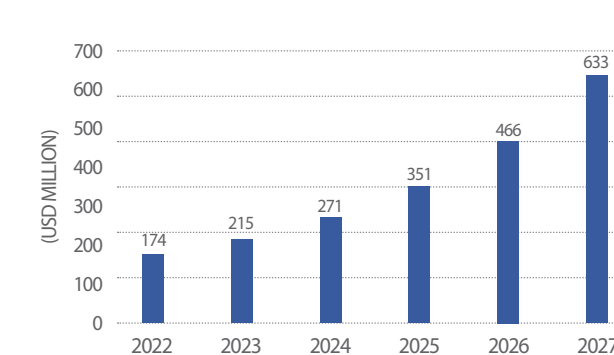
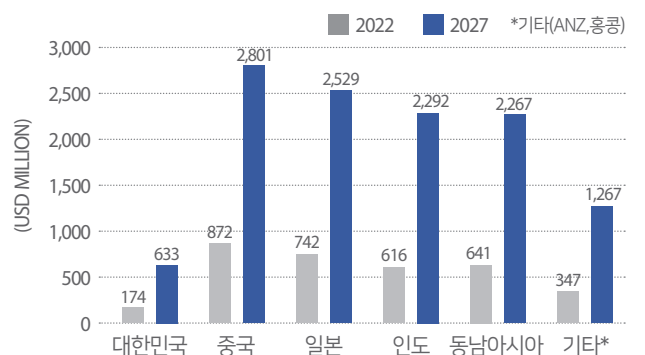


그림 5 자연어 처리의 국내 시장 규모 및 전망(2022~2027)



출처 : Markets&Markets, Natural Language Processing Market, 2022, (KISTI 재작성)

그림 6 자연어 처리의 아시아 지역별 시장 규모 및 전망(2022~2027)



세계 자연어 처리 시장을 지역별로 살펴보면, 북미 지역은 2022년 기준 세계 시장의 33.6%를 차지해 시장점유율이 가장 큰 것으로 나타났다. 그 이유는 미국의 독보적인 AI 기술력과 AI 분야에 대한 선제적인 투자 때문이라고 할 수 있다. 북미 지역은 2022년 52억 8,800만 달러에서 연평균 24.1%로 성장해 2027년 155억4,000만 달러에 이를 것으로 전망된다. 유럽 지역은 2022년 40억4,900만 달러에서 연평균 24.8%로 성장해 2027년 122억5,800만 달러에 이를 것으로 전망된다. 아시아태평양 지역은 2022년 33억9,300만 달

려에서 연평균 28.3%로 성장해 2027년 117억9,100만 달러에 달해 지역별 시장 규모 중에서 시장성장률이 가장 큰 것으로 예측된다. 중남미 지역의 자연어 처리 시장은 예측 기간 동안 연평균 25.5%로 성장할 것으로 예상되며, 중동 및 아프리카 지역은 예측 기간 동안 27.4%로 성장할 것으로 예상된다. 글로벌 차원에서 자연어 처리 시장의 증가 추세는 전 세계 산업 및 기업에서의 AI 서비스 개발 사례가 증가하고, 일상 속에서도 사용자의 편의성과 생산성 향상을 위해 AI 적용 서비스가 보급되고, 활용 분야의 확대에 영향을 받기 때문이다.

표 1

자연어 처리의 지역별 시장 규모 및 전망(2022~2027)

(단위: 백만달러)

구 분	2022	2023	2024	2025	2026	2027	CAGR(%)
북미	5,288	6,248	7,541	9,326	11,856	15,540	24.1
유럽	4,049	4,813	5,845	7,273	9,302	12,258	24.8
아시아태평양	3,393	4,150	5,187	6,643	8,744	11,791	28.3
중동/아프리카	1,730	2,099	2,602	3,304	4,314	5,806	27.4
중남미	1,289	1,542	1,880	2,342	2,992	4,018	25.5
계	15,750	18,852	23,055	28,888	37,208	49,413	25.7

출처 : Markets&Markets, Natural Language Processing Market, 2022

적용 분야별 시장 규모

자연어 처리의 적용 분야는 대화형 음성 응답(Interactive Voice Response, IVR), 문자 인식(Optical Character Recognition, OCR), 텍스트 분석, 음성 분석, 오토 코딩, 이미지 인식, 시물레이션&모델링 등의 세부 기술로 구분할 수 있다. 그 중에서 IVR이 가장 높은 시장점유율을 보이며, 2022년 31억7,900만 달러에서 연평균

26.4 %로 성장해 2027년 102억6,800만 달러에 이를 것으로 전망 된다. IVR이 높은 비중을 차지한 이유는 언택트 스토어, 금융권 컨택트센터, 상담 및 조사가 필요한 공공·의료 서비스 분야에 대화형 음성 응답 솔루션의 적용으로 편리한 언택트 소비 경험이 확산되었기 때문이며 독거 노인 돌봄, 장애인 일상 지원 등 사회적 약자의 불편을 해소하기 위한 AI 제품의 보급에 자연어 처리 기술을 적용한 사례가 증가하고 있어 향후 시장은 더욱 성장할 것으로 예상된다.

표 2

자연어 처리 적용 분야별 시장 규모 및 전망(2022~2027)

(단위: 백만달러)

구 분	2022	2023	2024	2025	2026	2027	CAGR(%)
대화형 음성 응답	3,179	3,827	4,708	5,934	7,687	10,268	26.4
문자 인식(OCR)	2,355	2,784	3,362	4,160	5,290	6,919	24.1
텍스트 분석	2,788	3,315	4,028	5,013	6,415	8,457	24.8
음성 분석	2,083	2,486	3,033	3,790	4,868	6,448	25.4
오토 코딩	1,446	1,744	2,151	2,717	3,528	4,724	26.7
이미지 인식	1,882	2,281	2,824	3,583	4,673	6,270	27.2
시물레이션 및 모델링	2,017	2,414	2,950	3,691	4,746	6,326	25.7
계	15,750	18,851	23,056	28,888	37,207	49,412	25.7

출처 : Markets&Markets, Natural Language Processing Market, 2022

표 3

자연어 처리 적용 분야별 활용 사례

구 분	정의	활용 사례
대화형 음성 응답	사전 녹음된 메시지 또는 텍스트-음성 기술 기반 음성 응답 인터페이스	고객 센터(상담, 예약 등 단순·반복 업무 응대), 대화형 AI 서비스 등
문자 인식(OCR)	문서 내 문자의 영상을 이미지 스캐너로 획득해 기계가 읽을 수 있는 문자로 변환	공공/의료/금융 서비스 내 스캔 서류 자동 판독, 이미지 번역 서비스 등
텍스트 분석	텍스트 데이터 내 정보 분석 및 추출	정보 검색, 질의 응답, 문서분 류, 문서 요약, 감정 분석, 리뷰 분석 등
음성 분석	음성을 인식해 텍스트로 변환(전사)하는 기술	회의록 작성 등 전사(transcript) 서비스, 음성 인식 서비스
오토 코딩	프로그래밍 소스 코드 생성	프로그래밍 교육, 개발자의 보조 도구 (ex. GPT-J, GPT-NeoX)
이미지 인식	인식된 이미지를 텍스트로 변환 및 설명	자동 자막 생성, 이미지 캡셔닝(ex. DALL-E)
시물레이션 및 모델링	자연어 이해를 기반으로 지식을 자가 학습하고, 인간의 사고 과정을 시물레이션 (학습, 추론)	인지 컴퓨팅, 디지털 휴먼, 자연어 생성

경쟁 현황

자연어 처리 기반 AI 산업은 글로벌 빅테크 기업이 대규모 언어 모델을 발표하고, 비즈니스 모델을 개발하는 형태로 경쟁이 가속화되고 있다. 대표적으로 OpenAI가 GPT-3 언어 모델 발표 이후 구글, 메타 등 글로벌 빅테크 기업에서는 대규모 언어 모델을 연이어 개발하였으며, 영어권 이외 국가에서도 AI 주권 확보를 위해 자국어 언어 모델의 자체 개발에 힘쓰고 있다. 국내에서도 네이버, 카카오, LG전자, SKT, KT 등

에서 적극적으로 대규모 언어 모델을 개발하고 있는 이유이다. 대규모 언어 모델은 방대한 데이터의 학습과 수 억, 수 조 단위의 파라미터³⁾ 연산을 위한 막대한 컴퓨팅 파워가 요구되기 때문에 압도적 자본력과 기술력을 바탕으로 '규모의 경제'를 확보한 글로벌 빅테크 기업이 선도하고 있으며, 소규모 기업의 경우 빅테크 기업이 확보한 인프라와 기술 등 자원을 활용해 생산성 강화 및 공공성과 연계한 특화 분야 AI 서비스 개발에 집중해 시장 경쟁력을 확보해 나갈 것으로 전망된다.

표 4

대용량 언어 모델 공개 현황

주요 기업		모델/서비스	적용 분야
해외	OpenAI	GPT-3	1,750억 개 파라미터, 언어 모델/언어 생성, 이해, 추론, 코딩, 오류 수정
		ChatGPT	1,750억 개 파라미터, “대화형” 언어 모델/GPT-3.5(사람의 피드백 학습에 반영)모델 기반/콘텐츠 생성(텍스트 생성, 번역, 챗봇, 오토코딩 등)
		GPT-4	1조 개 이상 파라미터 추정/GPT-3 성능 고도화 이미지 인식/분석 가능한 멀티 모달 모델/ChatGPT 유료 버전 적용


3) 파라미터(parameter): 인간 뇌의 뉴런 및 뉴런 간의 연결에 해당하는 매개 변수를 의미하며, 딥러닝에서 파라미터 개수가 많을수록 더 많은 정보를 저장 및 처리 가능한 고성능을 의미

주요 기업		모델/서비스	적용 분야
해외	구글	PaLM	5,400억 개 파라미터, 언어 모델/언어 생성, 이해, 추론, 코딩, 오류 수정
		람다(LaMDA)	1,370억 개 파라미터, “대화형” 언어 모델
		바드(Bard)	람다 모델 기반 대화형 AI 에이전트/최신 정보에 대한 답변 가능
	MS - Nvidia	MT-NLG	5,300억 개 파라미터, 언어 모델/컴퓨팅 환경(Nvidia A100), 병렬처 리(MS Deep-speed)/예측, 독해, 추론, 단어 중의성 해석
	메타	OPT-175B	1,750억 개 파라미터/비상업적 라이선스, 연구 목적으로 용도 제한
국내	네이버	HyperCLOVA	2,040억 개 파라미터, 언어 모델/한국어 최적화 모델 (한국어 97 %)/쇼핑,케어콜, 검색, 플레이 스 서비스 적용
	카카오브레인	KoGPT	300억 개 파라미터, 한국어 특화 모델/공정 부정 판단, 요약, 결론 예측, 질의 응답,, 다음 문장 생성
	LG AI 연구원	엑사원	3,000억 개 파라미터, 국내 최대 규모/언어, 이미지, 영상 등 멀티 모달 모델
	SKT	GLM	1,500억 개 파라미터/국립국어원 협력, 한국어 모델/챗봇, 텍스트 감성 예측, 응답 생성/에이닷 적용
		에이닷	GPT-3 모델 기반 대화형 AI 에이전트
	KT	믿음(MIDEUM)	2,000억 개 파라미터/대화형 AI 에이전트(상당 분야 특화)
	솔트룩스	루시아(Luxia)	한국어 대화형 언어 모델
		서치스튜디오	루시아 모델 기반 인지 검색(Cognitive Search) 서비스/개인화 서비스 제공

4) 분석자 인사이트

자연어 처리 기술은 대규모 언어 모델의 등장으로 인해 정확도가 견고해지면서 다양한 산업별 서비스가 점차 확대되어 AI 서비스 대중화를 예상해 볼 수 있다. 최근 OpenAI에서 공개한 ChatGPT를 통해서 알 수 있듯이 문서뿐만 아니라 음악, 영상, 이미지 등 다양한 콘텐츠를 생성하는 생성형 AI(Generative AI)가 주목받고 있다. 이처럼 AI 기술의 진화와 확장으로 인한 사용자의 혁신적인 경험은 AI 서비스에 대한 접근성을 높이면서 업무 프로세스의 자동화를 통한 비용 절감이라는 가시적 성과를 가져왔으며, 이러한 AI 기술의 창작 능력은 패션, 디지털 마케팅, 프로그래밍 등 사용자의 특정한 요구에 맞춘 콘텐츠 생성 관련 산업 분야에서도 변화 동인으로 작용할 것으로 전망된다.

따라서 AI 기술이 본격적으로 적용되면서 디지털 전환에 필수인 AI 반도체를 비롯해 자연어 처리 기술도 관련 시장의 성장과 함께 기

술의 고도화가 지속적으로 요구되는 등 자연어 처리 기술이 형성하는 시장의 변화를 예측하면서 관련 주요 영향 요인을 <표 5>와 같이 정리하였다. 

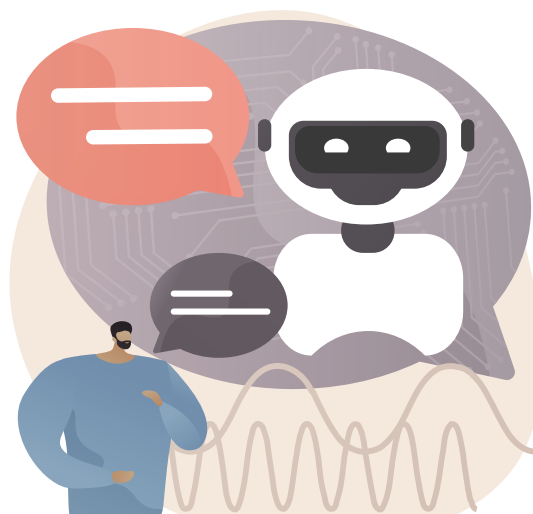


표 5 시장 영향 요인 분석

촉진 요인

- 대규모 언어 모델 등장으로 상용 서비스 품질 향상
- 인공지능 솔루션 인식 확대 및 관심 증가
- AI 컴퓨팅 파워 구축, AI 반도체 개발 등 초거대 AI 모델개발을 위한 후방 기술 성장

저해 요인

- 기술 개발에 대한 막대한 투자 비용으로 소규모 업체의 시장 진입 어려움
- 초거대 AI 도입 및 활용 역량 부족
- 대규모 모델에 대한 중속성(학습 데이터 접근이 어려움)

기회 요인

- 업무 자동화, 생산성 향상을 위한 기업의 AI 서비스 수요 증가
- 음성, 이미지 데이터와 접목한 멀티 모달 모델 확장, 생성형 AI 시장 확대
- 첨단 기술 제품에 대한 소비자 수용 및 선호도 증가
- 초거대 AI 모델의 실제 서비스 적용을 위한 경량화·최적화 기술 개발 필요

위험 요인

- 개인 정보 보호 및 보안에 대한 우려
- 생성된 콘텐츠의 저작권 문제, 유해 콘텐츠 생성 등 법적·윤리적 문제
- 학습 데이터의 편향성(성차별, 인종주의 등)
- 신뢰성(할루시네이션, 그럴듯한 답변 구사)

참고문헌

- [1] Natural Language Processing Markets, Markets and Markets, 2022.
- [2] 초거대 언어모델을 기반으로 한 AI 대화 인터페이스, 국어학회, 2023.
- [3] 실리콘밸리의 디지털 혁신 트렌드, THE AI REPORT, 2022. 12.
- [4] 2023년 글로벌 인공지능(AI) 기술개발 동향과 사업화 전략, CHO Alliance, 2023.
- [5] 초거대언어모델의 부상과 주요이슈, SPRI 이슈리포트(IS-158), 2023.





www.astinet.kr
에서 원문을 다운로드
받으실 수 있습니다.

ASTI MARKET INSIGHT



본원 (우)34141 대전광역시 유성구 대학로 245 한국과학기술정보연구원
T. 042) 869-1004, 1234 F. 042) 869-1091

분원 (우)02456 서울특별시 동대문구 회기로 66 한국과학기술정보연구원
T. 02) 3299-6114 F. 02) 3299-6244

