우리 마음속 첫번째 금융

우리금융그룹

기업금융·신성장 트렌드

금융업계 AI 확산으로 인한 리스크와 대응방안

2024. 5. 31.

- **생성형 AI기술이 급속도로 발전하며 금융업계에도 AI활용이 일반화되고 있는 가운데, 기술의 발전과 정부·금융당국의 규제만으로는 AI 관련 리스크를 완전히 대비할 수 없어 리스크를 세분화해 인식하고 대비할 필요**
  - 미국, EU, 한국 등 주요국은 AI 규제법률을 제정하거나 유관기관을 확정하고 보안 가이드라인을 발표하고 있으며, 주로 개인정보 보호와 차별금지 등에 집중
- **금융업과 연관성이 높은 리스크를 ①데이터 관련 리스크, ②산출물 관련 리스크, ③시장 변동성 확대 리스크의 3가지로 분류**
  - **(① 데이터 관련 리스크)** AI 시스템의 인풋이 되는 데이터 관리와 관련해 개인정보 유출 등 프라이버시 문제 뿐 아니라 유입 데이터가 보편·공정성을 잃는 편향(bias)이나 악의적으로 아웃풋을 조작하기 위한 데이터 오염(poisoning) 등의 위험이 높음
  - **(② 산출물 관련 리스크)** 인풋 데이터의 문제가 없더라도 생성형 AI의 내재적 특성상 주어진 질문에 대해 사실이 아니나 그럴듯한 답변을 제시하는 환각(Hallucination)현상이 발생해 잘못된 산출물(아웃풋)을 생성할 가능성
  - **(③ 시장 변동성 확대 리스크)** AI시스템의 인풋(데이터)과 아웃풋(산출물)의 문제가 없더라도 다수의 금융기관이 동일·유사한 AI모델을 활용해 금융시장에 참여하면 주어진 뉴스에 동일한 포지션을 취하는 투자 규모가 확대되어 버블·버스트 사이클이 심화될 수 있음
- **AI를 적극 도입해 업무혁신을 이루고 효율성을 향상시키는 동시에, AI의 확산이 초래할 수 있는 리스크를 인식, 이해하고 자체적인 대비·대응방안을 마련할 필요**
  - 내부 데이터의 유출, 악용방지 뿐 아니라 외부의 AI모델과 트레이닝 데이터를 활용할 때에도 금융기관의 목적과 상황에 적합하고 편향이나 오염의 가능성이 없는지 항시 점검
  - 생성형 AI를 중요한 업무에 높은 빈도로 투입할수록 결과물의 팩트체킹을 위한 인간의 점검빈도를 높이거나 별도의 외부 서비스를 활용하는 등 환각 현상에 대비
  - 빅데이터 활용과 적시성 등 AI·알고리즘 트레이딩의 장점을 수용해 시장 대응력을 개선하는 한편, AI 트레이딩의 보편화가 초래할 수 있는 시장의 동조화와 변동성 확대에 대비

작성자: 경영전략연구실 오태준 수석연구원 (02-2173-0560)

우리금융경영연구소

I. AI 활용도 확산과 관련 리스크 분석 필요성

■ 2022.12월 ChatGTP의 등장 이후 생성형 AI 기술이 급속도로 발전하며 금융업계에서도 업무의 전 영역에서 생산성 향상과 비용 절감을 위해 AI를 활용하기 시작

- 생성형 AI는 기존에 학습된 데이터를 바탕으로 새로운 콘텐츠를 생산해 낼 수 있어 금융 상품과 서비스 개발, 시장 분석과 투자 전략 수립, 고객 분석·분류와 마케팅, 고객 응대 등 업무의 전 영역에서 널리 적용 가능[1)]

  * JPMC는 연간 10억 달러를 투자해 금융정보 분석서비스 IndexGPT를 개발하고, Morgan Stanley와 Netwest는 AI 가상비서를 통해 고객별 맞춤형 서비스를 제공하는 등 글로벌 금융회사들의 AI 활용이 확대

■ 최근 들어 AI의 활용도 확산에 따른 리스크에 대한 우려가 제기되고 있으며, 주요국은 AI 기술에 대한 통제력 확보와 보안 강화 정책을 수립 중

- IMF(2023)는 생성형 AI가 금융업에서 개인정보 유출, 잘못된 의사결정, 시장 교란 등을 통해 금융시장의 신뢰성과 안전성에 실질적 위험을 미칠 수 있다고 지적[2)]

  - AI 기술이 급속도로 발전하면서 금융서비스에 확산되는 속도가 빨라 정부·금융당국이 필요한 규제와 가이드라인을 적시에 완비하기 힘들어, 공공·민간 금융기관 및 이해당사자들과 소통, 협력하면서 리스크에 대비할 필요성을 제기

- 미국, EU, 한국 등은 AI 규제 법률을 제정하거나 유관 기관을 확정하고 보안 가이드라인을 발표하고 있으며, 주로 개인정보 보호와 차별금지 등에 집중([붙임 1] 참조)

  - 미국의 소비자금융보호국(CFPB)는 금융업계에서 AI를 활용하는 과정에서 데이터 유출 등 보안 위험이나 잘못된 정보가 생성·유포되는 등의 리스크에 대해 모니터링

  - EU가 2024.3월 제정한 AI법은 생성형 AI가 HR 등에 활용될 경우의 차별금지 등에 대한 구체적인 조항을 포함

1) 우리금융경영연구소, “글로벌 금융회사의 생성형AI 활용 사례와 시사점”, (2024.02.)
2) IMF, “Generative Artificial Intelligence in Finance: Risk Considerations”, (2023.08)

* 금융감독원은 '금융분야 AI가이드라인'(2021)에서, 금융산업의 책임성, 데이터의 정확성과 안전성, 서비스의 투명성과 공정성, 소비자 권리 보장 원칙을 발표
* '금융분야 AI활용 활성화 및 신뢰확보방안'(2022)에서는 데이터 확보지원과 AI 제도 점검, AI 검증체계구축 방안을 추가 발표

※ 올 5월 한국과 G7 국가들, 호주와 UN, OECD, EU 대표와 각국의 글로벌 기업들이 참가한 'AI 서울 서밋'에서는 AI의 개발과 적용에서 안전과 혁신, 포용의 원칙을 확립하고 주요국들이 합동으로 협력해나갈 것을 공표

■ **추후 AI 기술이 발전하고 규제가 완비되어도, 생성형 AI와 금융업의 특성상 모든 종류의 리스크에 완벽하게 대비할 수 없어 리스크를 세분화하여 분석하고 대응방안을 마련할 필요**

- 인프라 장애나 서비스 운영 미숙, 인적오류 등으로 인한 Tech Risk는 이미 발생사례가 다수 보고[3]되고 있는 반면, 생성형 AI의 금융서비스 적용은 아직 도입 초기단계로 리스크가 현실화되지 않았고 관련 규제도 정리되지 않았음

II. 금융업계 AI 도입 관련 리스크의 성격과 실현시 파급효과

연구소는 생성형 AI의 도입과 AI 활용 확대에 관련된 리스크 중 금융업과 연관성이 높은 리스크를 ①데이터 관련 리스크, ②산출물 관련 리스크, ③시장 변동성 확대 리스크의 3가지로 분류

① 데이터 관련 리스크

■ **AI 시스템의 인풋이 되는 데이터 관리와 관련해 개인정보 유출 등 프라이버시 문제뿐 아니라 유입 데이터의 편향(bias)이나 오염(poisoning)등의 위험도가 높을 수 있음**

- 개개인의 데이터가 익명화되어서 취급·저장된다 해도 데이터가 담고 있는 정보를 통해 해당 개인의 신원을 역추적해 특정이 가능할 수 있어 데이터 유출 방지가 중요
- 트레이닝 데이터의 내용이나, 데이터를 프로세스하는 알고리즘, 아웃풋을 해석하고 적용하는 인식과정에서 보편·공정성이 훼손될 가능성이 있음(데이터 편향)[4]

3) 우리금융경영연구소, "경영 Insight, 금융회사의 Tech Risk 사례와 시사점", (2024.02.02.)
4) National Institute of Standards and Technology, "There's more to AI bias than biased data", (2022.03.16.)

- 온라인 마케팅에 보편화된 검색엔진최적화(SEO, Search Engine Optimization) 테크닉이 AI 트레이닝 데이터에 영향을 주기 위해 사용될 가능성이 높으며, 이에 따라 데이터 편향 위험이 상승(IMF)[5]

• 시스템과 보안의 취약부문을 악용해 AI 모델을 무력화하거나 아웃풋을 조작하기 위해 외부에서 편향된 데이터를 주입하는 데이터 오염(poisoning)의 가능성도 존재[6]

- 이용되고 있는 AI 모델의 구조와 사이버보안 프로토콜에 대해 알고 있는 인사이더(조직 내부인이나 정보를 입수한 개인/단체)가 데이터 오염을 시도할 경우 공격이 수월하게 이루어질 수 있고 피해가 큼

• 금융기관의 의사결정은 개인이나 기업고객의 재무에 결정적인 영향을 미칠 수 있으며, 금융시장을 교란시키는 파급효과를 초래할 수도 있어 이해당사자들이 데이터 처리 과정에 개입해 결과를 조정하려는 인센티브가 존재

※ AI 모델에 주입된 데이터와 아웃풋의 정확한 인과관계를 설명할 수 없는 AI의 특성(lack of explainability)도 데이터 편향이나 데이터 오염에 대한 정확한 진단과 교정을 어렵게 하는 요인

② 산출물 관련 리스크

■ **인풋 데이터의 문제가 없더라도 생성형 AI의 내재적 특징에서 비롯된 환각(Hallucination) 현상으로 잘못된 산출물(아웃풋)이 발생할 위험이 있음**

• 환각 현상은 주어진 질문에 대해 사실이 아니나 그럴듯한 답변을 꾸며내어 대답하는 것으로, 생성형 AI를 활용하는 산업계에서 실질적인 피해와 신뢰저하 사례가 발생

* 2024.2월, 캐나다 행정재판소는 Air Canada의 AI챗봇이 사실과 다른 내용을 고객에게 안내한 부분에 대해 기업이 책임을 지고 고객에게 환불해야 한다고 결정[7]

* 2023.6월 미국의 맨해튼 지방법원은 뉴욕주 변호사 2명이 ChatGPT를 이용해 작성한 법원 제출 서류에서 6건의 사실이 아닌 거짓 판례를 인용한 것에 대해 5천 달러의 벌금을 부과[8]

5) IMF, “Generative Artificial Intelligence in Finance: Risk Considerations”, (2023.08)
6) Crowdstrike, “Data poisoning: the exploitation of generative AI”, (2024.03.20.)
7) Wired, “Air Canada has to honor a refund policy its chatbot made up”, (2024.02.17.)
8) Reuters, “New York lawyers sanctioned for using fake ChatGPT cases in legal brief”, (2023.06.26.)

- 생성형 AI는 작동 원리상 ①학습을 바탕으로 새로운 결과를 생성하고, ②대규모 데이터를 압축해 프로세스하며, ③자연스러운 답변 생성을 위해 결과를 취사선택하는 성질이 있어 근본적으로 환각을 완전히 제거할 수 없음[9)]
  - 양질의 데이터를 대량으로 학습한다 해도 항상 기존의 데이터와 일치하지 않는 질문이 발생하며, 저장공간의 한계로 데이터를 압축·복구하는 과정에서 유실이 발생하고 팩트와 유사하나 거짓된 답변 생성이 불가피
- 질문이 복잡할수록, 기존에 많이 논의되지 않은 내용일수록 환각이 발생할 확률이 높으며, 2024.5월 현재 생성형 AI들이 환각으로 잘못된 산출물을 생산할 확률은 평균 2.5% ~ 22%로 나타남[10)]
  - OpenAI의 최신 프리미엄 서비스인 GPT 4 Turbo의 환각비율이 2.5%로 가장 낮으며 GTP 4가 3.0%, Microsoft의 Orca-2-13b는 3.2%, Meta가 제공하는 Llama 3 70B가 4.5% 등([붙임 2] 참조)
- 금융시장에 구조적인 변화나 예기치 못한 상황이 발생할 경우 환각 현상으로 생성형 AI가 잘못된 의사결정을 내려 상당한 피해를 미치고 금융기관의 신뢰가 훼손될 가능성 존재[11)]

③ 시장 변동성 확대 리스크

■ **AI 시스템의 인풋(데이터)이나 아웃풋(산출물) 모두 문제가 없을 경우에도, 다수의 투자자가 동일·유사한 AI 모델을 활용해 금융시장에 참여하면 군집행동 (Herding behavior)이 강화되어 버블-버스트 사이클이 심화될 우려**

- 알고리즘 트레이딩은 미국 주식 거래의 60~73%, 유럽의 60%, 아시아의 45%를 차지하고있고 생성형 AI의 발전으로 금융기관뿐 아니라 일반 투자자들에까지 급격히 확산[12)]
  - AI 모델의 발전으로 기존의 단순한 예측형 알고리즘에서 벗어나 더 많은 데이터를 활용해 실시간으로 고도화된 분석을 수행할 수 있으며, 일반인들도 생성형 AI를 이용해 알고리즘 제작이 가능

9) Scientific America, “AI Chatbots will never stop hallucinating”, (2024.04.05.)
10) Vectara, “Hallucination Leaderboard”, (2024.05.14.)
11) IMF, “Generative Artificial Intelligence in Finance: Risk Considerations”, (2023.08)
12) Benzinga, “What percentage of stock trades are made by bots and algorithms?”, (2023.06.14.)

- AI를 활용한 트레이딩 알고리즘은 주어진 경제·금융환경과 데이터를 바탕으로 빠른 속도로 각종 자산의 매수·매도 포지션을 결정해 동일 AI를 이용하는 다수의 투자자가 일시에 같은 결정을 내리면 시장 변동성이 급격히 확대[13)]

  ※ AI 모델도 인터넷 검색이나 웹 커머스와 같이 이용자가 많을수록 서비스의 질이 개선되는 네트워크 효과(network-effect)가 있어, 최종적으로 가장 우수한 소수의 서비스를 대다수 고객이 이용할 가능성이 높음

- 시장의 시스템 리스크가 증가하는 것으로 AI·알고리즘 트레이딩을 이용하지 않는 투자자나 금융기관도 위험에 노출

III. 시사점

■ **금융회사들은 AI를 적극 도입해 업무 혁신을 이루고 효율성을 향상시키는 동시에, AI의 확산이 초래할 수 있는 리스크를 인식, 이해하고 자체적인 대비·대응방안을 마련할 필요**

- 내부 데이터의 유출·악용 방지뿐 아니라 금융회사가 활용하고 있는 외부의 AI 모델과 트레이닝 데이터가 내부의 목적과 상황에 적합하고 편향이나 오염의 가능성이 없는지 항시 점검

  - 임직원이나 업무상 협력 관계로 내부 시스템에 접근하기 용이한 인물·조직의 부주의나 고의로 데이터의 유출·편향·오염이 발생하지 않는지 모니터링을 강화

- 생성형 AI를 중요한 업무에 높은 빈도로 투입할수록 결과물의 팩트체킹을 위한 인간의 점검빈도를 높이거나 별도의 외부 서비스를 활용하는 등 환각 현상에 대비

  * 엔비디아는 2023.4월, AI챗봇이 환각으로 인해 잘못된 정보를 제공하거나 바람직하지 않은 아웃풋을 생산하는 것을 방지하고 보안을 강화하기 위한 소프트웨어 서비스 'NeMo Guardrails'를 발표[14)]

- 빅데이터 활용과 적시성 등 AI·알고리즘 트레이딩의 장점을 수용하여 시장 대응력을 개선하는 한편, AI 트레이딩의 보편화가 초래할 수 있는 시장의 동조화와 변동성 확대에도 대비

13) Bloomberg, "AI risks amplifying 'Herd-Like' behavior in trading, BOE says", (2024.05.07.)
14) CNBC, "Nvidia has a new way to prevent A.I. chatbots from 'hallucinating' wrong facts", (2023.04.25.)

붙임 1 국내·외 AI 보안·리스크 관리 관련 주요 정책

■ **미국: AI 전반을 규제하는 단일 법률이나 독자기관은 아직 없으며, 다수의 유관 기관과 소비자 보호 법률 등을 통해 AI 관련 보안 리스크 등을 규제[15)]**

- 2023.3월, 연방거래위원회(FTC), 평등고용기회위원회(EEOC), 소비자금융보호국(CFPB)과 법무부(DJ)는 각 기관의 유관 영역에서 'AI를 포함한 소프트웨어 및 알고리즘 프로세스'에 대해 관리·감독 권한이 있다는 공동 성명을 발표
- 2023.10월 바이든 정부는 AI개발과 활용에 대한 대통령령(Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence)을 통해 8가지 원칙을 발표
  1. AI는 안전하고 보안이 유지되어야 함
  2. AI 선도를 위해 미국은 책임감 있는 혁신과 경쟁, 협력을 장려해야 함
  3. 책임있는 AI개발과 사용은 미국 근로자를 지원방식으로 이루어져야 함
  4. AI정책은 형평성과 시민권을 증진해야 함
  5. 일상생활에서 AI와 AI 지원제품을 사용하고 상호작용하고 구매하는 미국 시민들의 이익이 보호되어야 함
  6. 개인정보와 시민의 자유가 보호되어야 함
  7. 연방정부는 AI 사용에 따른 위험을 관리해야 함
  8. 연방정부는 사회와 경제, 기술 발전에서 글로벌 리더십을 행사해야 함

■ **EU: 2024.3월 AI법(Artificial Intellegence Act)을 제정해 종합적으로 AI 기술의 개발과 활용 전반에 대해 관리·감독하기 위한 기반을 마련**

- AI 시스템을 활용 데이터 민감도와 방식에 따라 4개의 리스크 계층으로 분류하고, 위험도가 높은 AI('금지'와 '고위험'으로 분류된 AI)에 대해서는 더욱 엄격한 조치를 적용
- 생체인식분류, 사회적 행동이나 개인적 특성을 기반으로 개개인을 평가·분류하는 AI는 '금지'

15) White&Case, "AI Watch: Global regulatory tracker - United States", (2024.05.13.)

- 고용 및 HR 시스템에 사용되는 등의 AI 시스템은 '고위험'으로 분류해 엄격한 관리·감독을 수행
- 유럽연합 집행위원회와 AI 오피스가 법률 시행 18개월 내에 실질적, 구체적인 AI 시스템 분류 지침을 개발하며, 법률의 집행은 EU 회원국별 감시 당국과 EU의 AI 사무국(2024.2월 설립)이 담당
- '금지된' AI 위반에는 전세계 연 매출액의 7%나 3,500만 유로 중 큰 금액의 벌금이, '고위험' AI 시스템의 요구사항을 위반한 경우에는 전 세계 연 매출액의 3%, 부정확하거나 불안하거나 오해의 소지가 있는 정보를 제공한 경우에는 전 세계 연 매출액의 1.5%의 벌금이 부과

■ **한국: 국내 AI 산업의 전반적인 진흥, 규제 방안을 담은 'AI기본법'의 제정이 지연되고 있는 가운데, 금융감독원은 금융분야의 AI 서비스에 대한 신뢰 향상을 위한 가이드라인을 제정(2021,2022년)**

- 수년간 여야의원들이 발의한 인공지능관련 7개 법안을 통합한 '인공지능 산업 육성 및 신뢰 기반 조성에 관한 법률안'(AI 기본법)이 2023.2월 국회 과학기술정보방송통신위원회를 통과했으나 1년 이상 국회에 계류
  - 국가 인권위원회와 시민단체 등이 현 AI 기본법에 명시된 AI 기술의 '우선 허용, 사후 규제'관련 조항에 반대하고 안전과 인권보호를 더욱 강조하는 방향으로 수정해야 한다고 주장
- 금융위는 2021.7월 '금융분야 AI 가이드라인'과 2022.8월 '금융권 인공지능 활용 활성화 및 신뢰확보 방안'을 발표

금융분야 AI 가이드라인(2021)[16]

핵심 가치	내용
금융산업의 책임성 강조 (3중 내부통제장치 마련)	- (AI 윤리) AI 서비스 개발, 운영시 준수해야할 원칙, 기준 수립 - (AI 조직) AI의 잠재위험을 평가, 관리 할 구성원의 역할, 책임, 권한을 서비스 전 단계(기획, 설계, 운영, 모니터링)에 걸쳐 구체적으로 정의 - (위험관리) AI 서비스 자체 평가, 관리정책 마련
데이터의 정확성, 안전성 확보	- AI 학습 데이터의 품질 검증, 개선 및 최신성 유지 - 개인정보 보호(오, 남용 방지, 불필요한 개인신용정보 처리 최소화 등)
서비스의 투명성, 공정성 담보	- 서비스 특성에 맞는 위험 통제 - 소비자 차별 방지 등 서비스 특성별 공정성 기준을 설정, 평가
금융소비자 권리의 보장	- 금융소비자 앞 AI이용 사전고지, 소비자의 권리 및 이의신청, 민원제기 등 권리구제 방안 등을 알기 쉽게 안내

자료: 금융위원회, KDB미래전략연구소, 우리금융경영연구소

금융분야 AI 활용 활성화 및 신뢰확보 방안(2022)

핵심 가치	내용
양질의 빅데이터 확보 지원	- 가명정보 재사용을 허용하는「금융 AI 데이터 라이브러리」구축 - 협업을 통한 데이터 공동 확보 - 데이터 전문기관 추가 지정
AI활성화를 위한 제도 점검	- AI개발, 활용 안내서 발간 - 설명 가능한 AI 요건 마련 - 망분리 및 클라우드 규제 개선
신뢰받는 AI 활용환경 구축	- 금융 AI 테스트베드 구축 - AI 기반 신용평가모형 검증체계마련 - AI 보안성 검증체계 구축

자료: 금융위원회, 우리금융경영연구소

16) KDB 미래전략연구소, '금융분야 AI 가이드라인' 및 금융권의 대응', (2021.10.12.)

붙임 2 주요 생성형 AI의 평균 환각 확률(2024.5월)[17]

Model	Hallucination Rate	Factual Consistency Rate	Answer Rate	Average Summary Length (Words)
GPT 4 Turbo	2.5 %	97.5 %	100.0 %	86.2
Snowflake Arctic	2.6 %	97.4 %	100.0 %	68.7
Intel Neural Chat 7B	2.8 %	97.2 %	89.5 %	57.6
GPT 4	3.0 %	97.0 %	100.0 %	81.1
Microsoft Orca-2-13b	3.2 %	96.8 %	100.0 %	66.2
GPT 3.5 Turbo	3.5 %	96.5 %	99.6 %	84.1
GPT 4o	3.7 %	96.3 %	100.0 %	77.8
Cohere Command R Plus	3.8 %	96.2 %	100.0 %	71.2
Mixtral 8x22B	3.8 %	96.2 %	99.9 %	92.0
Cohere Command R	3.9 %	96.1 %	99.9 %	51.2
Microsoft Phi-3-mini-128k	4.1 %	95.9 %	100.0 %	60.1
Mistral 7B Instruct-v0.2	4.5 %	95.5 %	100.0 %	106.1
Llama 3 70B	4.5 %	95.5 %	99.2 %	68.5
Google Gemini 1.5 Pro	4.6 %	95.4 %	89.3 %	82.1
Google Gemini Pro	4.8 %	95.2 %	98.4 %	89.5
Microsoft WizardLM-2-8x22B	5.0 %	95.0 %	99.9 %	140.8
Microsoft Phi-3-mini-4k	5.1 %	94.9 %	100.0 %	86.8
Llama 2 70B	5.1 %	94.9 %	99.9 %	84.9
Google Gemini 1.5 Flash	5.3 %	94.7 %	98.1 %	62.8
Llama 3 8B	5.4 %	94.6 %	99.8 %	79.7
Llama 2 7B	5.6 %	94.4 %	99.6 %	119.9
Llama 2 13B	5.9 %	94.1 %	99.8 %	82.1

17) Vectara, “Hallucination Leaderboard”, (2024.05.14.)

Anthropic Claude 3 Sonnet	6.0 %	94.0 %	100.0 %	108.5
Databricks DBRX Instruct	6.1 %	93.9 %	100.0 %	85.9
Google Gemma-1.1-7b-it	6.3 %	93.7 %	100.0 %	64.3
Anthropic Claude 3 Opus	7.4 %	92.6 %	95.5 %	92.1
Google Gemma-7b-it	7.5 %	92.5 %	100.0 %	113.0
Cohere-Chat	7.5 %	92.5 %	98.0 %	74.4
Cohere	8.5 %	91.5 %	99.8 %	59.8
Anthropic Claude 2	8.5 %	91.5 %	99.3 %	87.5
Microsoft Phi 2	8.5 %	91.5 %	91.5 %	80.8
Google Palm 2	8.6 %	91.4 %	99.8 %	86.6
Mixtral 8x7B	9.3 %	90.7 %	99.9 %	90.7
Amazon Titan Express	9.4 %	90.6 %	99.5 %	98.4
Mistral 7B Instruct-v0.1	9.4 %	90.6 %	98.7 %	96.1
Google Palm 2 Chat	10.0 %	90.0 %	100.0 %	66.2
Google Gemma-1.1-2b-it	11.2 %	88.8 %	100.0 %	66.8
Google flan-t5-large	15.8 %	84.2 %	99.3 %	20.9
tiiuae falcon-7b-instruct	16.2 %	83.8 %	90.0 %	75.5
Apple OpenELM-3B-Instruct	22.4 %	77.6 %	99.3 %	47.2