AI 윤리 리스크 진단 보고서

진단 대상: perplexity
진단 일시: 2025-10-23 15:57

Ⅰ. 서비스 개요

서비스 유형: 생성형 AI

주요 목적: Perplexity AI 요약

기술적 구조:
Perplexity AI는 검색 엔진과 생성형 AI를 결합한 차세대 AI 어시스턴트입니다. 이 시스템은 대규모 언어 모델(LLM)과 자연어 처리(NLP)를 활용하여 사용자의 질문에 대해 실시간으로 웹에서 정보를 검색하고, 이를 요약하여 제공하는 구조로 되어 있습니다. 각 답변에는 출처 링크가 포함되어 있어 정보의 신뢰성을 높입니다.

서비스 목적:
Perplexity AI는 비즈니스, 학술 연구, 개인 정보 수집 등 다양한 분야에서 신뢰할 수 있는 정보를 제공하여 사용자의 리서치 효율성을 높이는 것을 목표로 합니다. 특히 최신 정보가 필요한 상황에서 유용하며, 사용자가 쉽게 접근하고 활용할 수 있도록 설계되었습니다.

주요 기능:
1. 실시간 정보 검색: 사용자가 입력한 질문에 대해 웹에서 즉시 정보를 수집하고 요약하여 제공.
2. 출처 명시: 각 답변에 출처 링크를 명확히 표기하여 정보의 신뢰성을 검증할 수 있도록 함.
3. 검색 범위 조정: 특정 도메인이나 장르로 검색 범위를 좁힐 수 있어 목적에 맞는 정보 수집 가능.
4. 모바일 최적화: PC와 스마트폰 모두에서 쉽게 접근할 수 있도록 설계되어 이동 중에도 사용 가능.
5. 문서 업로드 및 요약: PDF나 텍스트 파일을 업로드하여 그 내용을 바탕으로 질문할 수 있는 기능 제공.

Perplexity AI는 무료 및 유료 플랜을 제공하며, 특히 유료 플랜에서는 고급 모델을 선택할 수 있어 더 복잡한 리서치나 창작 작업에 적합합니다.

Ⅱ. 초기 윤리 리스크 평가

공정성 (Fairness): 4.0점
공정성 (Fairness): 4/5

편향성 (Bias): 3.0점
편향성 (Bias): 3/5

투명성 (Transparency): 4.0점
투명성 (Transparency): 4/5

설명가능성 (Explainability): 3.0점
설명가능성 (Explainability): 3/5

프라이버시 (Privacy): 5.0점
프라이버시 (Privacy): 5/5

Ⅲ. 사용자 피드백

hallucination, 속도

Ⅳ. 피드백 반영 후 재평가 결과

공정성 (Fairness): 4.0점
공정성 (Fairness): 4점

편향성 (Bias): 3.0점
편향성 (Bias): 3점

투명성 (Transparency): 4.0점
투명성 (Transparency): 4점

설명가능성 (Explainability): 3.0점
설명가능성 (Explainability): 3점

프라이버시 (Privacy): 5.0점
프라이버시 (Privacy): 5점

Ⅴ. 최종 개선 권고안

AI 서비스의 윤리 리스크 평가 결과를 바탕으로 각 항목에 대한 구체적인 개선 권고안을 제시하겠습니다. 이 권고안은 EU, OECD, UNESCO의 가이드라인과 연계하여 설명하겠습니다.

1. 공정성 (Fairness) - Score: 4.0
개선 권고안:
• 다양한 데이터 소스 활용: AI 모델의 학습에 사용되는 데이터셋이 다양한 인구 집단을 포함하도록 보장해야 합니다. 이를 통해 특정 집단에 대한 차별을 방지할 수 있습니다.
• 정기적인 공정성 평가: AI 시스템의 결과물이 공정한지 정기적으로 평가하고, 필요시 조치를 취하는 프로세스를 마련해야 합니다.
• EU AI Act 준수: EU의 AI 법안에 따라 공정성을 보장하는 절차를 마련하고, 이를 통해 AI의 공정성을 지속적으로 모니터링해야 합니다.

2. 편향성 (Bias) - Score: 3.0
개선 권고안:
• 편향성 검토 및 수정: AI 시스템에서 발생할 수 있는 편향을 사전에 식별하고, 이를 수정하기 위한 알고리즘 및 프로세스를 도입해야 합니다.
• 다양한 이해관계자와의 협력: 다양한 배경을 가진 이해관계자와 협력하여 편향성을 줄이기 위한 피드백을 받을 수 있는 구조를 마련해야 합니다.
• OECD 가이드라인 활용: OECD의 인공지능 원칙을 참고하여 편향성을 줄이기 위한 윤리적 기준을 수립해야 합니다.

3. 투명성 (Transparency) - Score: 4.0
개선 권고안:
• AI 시스템의 결정 과정 문서화: AI의 결정 과정과 사용된 데이터에 대한 정보를 명확하게 문서화하고, 이를 사용자에게 제공해야 합니다.
• 사용자 교육: 사용자에게 AI 시스템의 작동 방식과 그 한계에 대해 교육하여 이해를 높이는 프로그램을 운영해야 합니다.

• UNESCO의 투명성 원칙 준수: UNESCO의 AI 윤리에 대한 원칙을 따르며, AI 시스템의 투명성을 높이기 위한 정책을 수립해야 합니다.

4. 설명가능성 (Explainability) - Score: 3.0
개선 권고안:
• 설명 가능한 AI 모델 개발: 사용자가 AI의 결정 과정을 이해할 수 있도록 설명 가능한 AI 모델을 개발해야 합니다.
• 사용자 피드백 반영: 사용자로부터 받은 피드백을 바탕으로 설명 가능성을 지속적으로 개선해야 합니다.
• EU AI Act의 설명가능성 요구사항 준수: EU AI 법안에서 요구하는 설명가능성 기준을 충족하기 위한 절차를 마련해야 합니다.

5. 프라이버시 (Privacy) - Score: 5.0
개선 권고안:
• 데이터 보호 정책 강화: 개인 데이터의 수집, 저장, 처리에 대한 강력한 데이터 보호 정책을 수립하고 이를 준수해야 합니다.
• 사용자 동의 절차 개선: 사용자로부터 명확한 동의를 받을 수 있는 절차를 마련하고, 사용자가 자신의 데이터에 대한 통제권을 가질 수 있도록 해야 합니다.
• GDPR 준수: 유럽 일반 데이터 보호 규정(GDPR)을 철저히 준수하여 개인 정보 보호를 강화해야 합니다.

이러한 개선 권고안들은 AI 서비스의 윤리적 리스크를 줄이고, 사용자와 사회에 대한 신뢰를 구축하는 데 기여할 것입니다. 각 권고안은 관련 국제 가이드라인과 연결되어 있어, 글로벌 스탠다드에 부합하는 방향으로 나아갈 수 있도록 돕습니다.

※ 본 보고서는 Human-in-the-loop 기반 AI 윤리 평가 결과입니다.