ODI Reliability White Paper (Beta)

Version : 1.0

Issue Date : 2025-06-10

Status : Beta

Owner : 국제AI교육원

본 문서는 국제AI교육원이 운영하는 교육용 AI "오디(ODI)"의 운영 원칙, Human Oversight, 신뢰성 및 활용 한계를 설명하기 위한 공개 문서입니다.

본 문서는 베타 운영 기준으로 작성되었으며 향후 운영 경험과 개선 활동에 따라 변경될 수 있습니다.

ODI Reliability White Paper (Beta)

1. Executive Summary
2. ODI란 무엇인가?

- Audit Thinking Coach
- Intended Use
- 적용 범위

3. 왜 AI 거버넌스를 공개하는가?

- 생성형 AI의 한계
- 신뢰성 문제
- Human Oversight 필요성

4. ODI 운영 원칙

- Intended Use
- Human Oversight
- 활용 범위 제한
- 공식 판단 대체 금지

5. Hallucination Risk Management

- 환각 위험 인식
- Evidence 우선 원칙
- 추가 확인 우선 원칙
- 불확실성 명시 원칙

6. Data Protection and Input Control

- 개인정보 입력 제한
- 기업기밀 입력 제한
- 교육용 사용 원칙

7. 현재 운영 중인 통제

- 정책 공개
- 사용자 안내
- 운영자 검토
- Knowledge 관리

8. 현재 한계

- 베타 운영
- 자동 환각 측정 부재
- 자동 피드백 분석 부재
- Human Review 의존

9. 향후 개선 방향

- Feedback Management
- Change Management
- Risk Register
- AI Governance Maturity

10. 42001 관점에서의 시사점

- AI 챗봇은 어떻게 심사할 것인가?
- Human Oversight Evidence는 무엇인가?
- 실제 심사 시 예상 질문

11. 결론

ODI는 ISO 심사 사고(Audit Thinking) 학습을 지원하기 위한 교육용 AI 시스템으로 운영되고 있습니다.

국제AI교육원은 생성형 AI의 가능성과 한계를 동시에 인식하며, Human Oversight, 활용 범위 제한, 데이터 입력 통제 및 환각(Hallucination) 위험 최소화 원칙을 적용하여 책임 있는 운영을 추구하고 있습니다.

현재 ODI는 베타 운영 단계에 있으며 모든 위험을 제거하거나 완전한 정확성을 보장하지 않습니다.

따라서 ODI의 답변은 학습 참고자료로 활용되어야 하며, 중요한 심사 판단과 의사결정은 반드시 사람의 검토를 거쳐야 합니다.

본 문서는 현재 운영 수준을 투명하게 공개하기 위한 것이며, 향후 운영 경험과 개선 활동에 따라 지속적으로 개정될 수 있습니다.

Document Control

Version History

v1.0 (2025-06-10)

- Initial Release