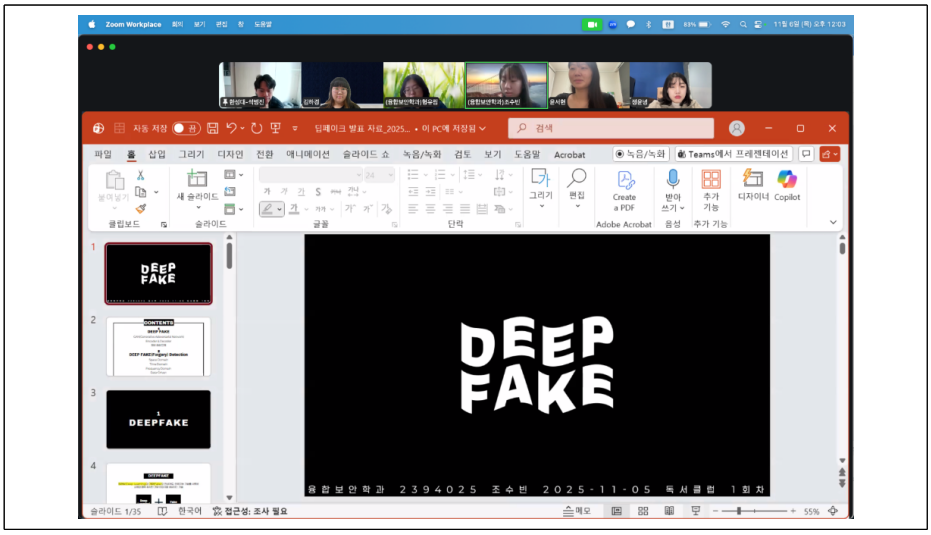
독서클럽 모임 보고서 - 석.아. (석병진과 아이들)

<table>
<tr><td rowspan="8">1주차</td><td>일시</td><td colspan="2">11월 6일 12 : 00 Zoom</td></tr>
<tr><td rowspan="6">참여 학생</td><td>클럽원 정보</td><td>참석 여부</td></tr>
<tr><td>김하경(2394010)</td><td>O</td></tr>
<tr><td>윤서현(2394048)</td><td>O</td></tr>
<tr><td>정윤녕(2394043)</td><td>O</td></tr>
<tr><td>조수빈(2394025)</td><td>O</td></tr>
<tr><td>형유림(2394016)</td><td>O</td></tr>
<tr><td>진도</td><td>도서명: 위험한 인공지능</td><td>진도페이지: p.173 ~ p.206</td></tr>
<tr><td>토론 내용</td><td colspan="2"> 

[좌측부터 석병진 교수님, 김하경, 형유림, 조수빈, 윤서현, 정윤녕]

김하경: 딥페이크가 사실 검증을 어렵게 만든다는 점에 주목했습니다. 특히 영상 조작이 여론 형성과 사회적 신뢰에 미치는 영향을 예로 들어 탐지 기술과 법적 규제가 함께 발전해야 한다는 의견을 제시했습니다.

윤서현: 딥페이크 제작 과정에서 발생하는 개인정보 무단 수집 문제를 강조했습니다. 온라인 콘텐츠 재사용이 범죄에 악용될 수 있어 법적 규제와 인식 개선이 필요하다고 이야기했습니다.

정윤녕: 생성형 AI를 이용한 피싱 자동화의 증가에 대해 논의했습니다. 문장 자연스러움이 향상되면서 기존 피싱 탐지 방법이 더 이상 충분하지 않다고 보았고, 사용자 교육 강화가 중요하다는 의견을 밝혔습니다.

조수빈: 5장 딥페이크를 발표했습니다. 기술이 고도화되며 정치적 조작 등 사회적 혼란으로 이어질 수 있다고 분석했고, 공격 대비 탐지 기술의 발전 속도가 더뎌 대응 전략 마련이 필요하다고 했습니다.

형유림: 딥페이크 탐지 기술의 한계와 인증 체계 보안에 대한 의문을 제기했습니다. 얼굴 기반 인증 방식이 무력화될 수 있어 대체 기술 연구가 필요하다고 의견을 남겼습니다.</td></tr>
</table>

<table>
<tr><td rowspan="8">2주차</td><td>일시</td><td colspan="2">11월 19일 19 : 00 코딩라운지 101호</td></tr>
<tr><td rowspan="6">참여 학생</td><td>클럽원 정보</td><td>참석 여부</td></tr>
<tr><td>김하경(2394010)</td><td>O</td></tr>
<tr><td>윤서현(2394048)</td><td>O</td></tr>
<tr><td>정윤녕(2394043)</td><td>O</td></tr>
<tr><td>조수빈(2394025)</td><td>O</td></tr>
<tr><td>형유림(2394016)</td><td>O</td></tr>
<tr><td>진도</td><td>도서명: 위험한 인공지능</td><td>진도페이지: p.30 ~ p.111</td></tr>
<tr><td>토론 내용</td><td colspan="2"> 

[좌측부터 석병진 교수님, 조수빈, 형유림, 정윤녕, 윤서현, 김하경]

김하경: 컴퓨터 비전 기술이 감시 체계를 자동화하며 오탐이나 오판이 발생할 경우 책임 소재가 모호해질 수 있다고 언급했습니다. AI 판단의 신뢰성 확보가 필수적이라는 의견을 제시했습니다.

윤서현: 적대적 공격이 실제 서비스 회피에 활용될 수 있다는 점을 우려했습니다. 작은 이미지 변조만으로도 분석이 실패할 수 있어 보안 기능 고도화가 필요하다고 말하며 토론에 참여했습니다.

정윤녕: 2장 컴퓨터 비전에 대해 발표하며 그중 사이버 보안과 비전 파트를 중심으로 설명했습니다. 영상 기반 탐지를 통해 보안 효율을 높일 수 있지만 개인정보 침해가 발생할 수 있어 관리가 필요하다고 했습니다.

조수빈: 적대적 공격이 실제 서비스에서 보안 회피에 활용될 가능성에 주목했습니다. 작은 이미지 변조만으로 시스템이 오작동한다면 사용자 피해가 커질 수 있다는 점을 지적했습니다.

형유림: 3장 적대적 학습을 발표하며 생성 AI와 공격 기술을 중점적으로 다뤘습니다. 백도어 삽입 등 모델 자체를 노리는 위협이 커지고 있어 적응형 방어 체계를 강화해야 한다고 강조했습니다.</td></tr>
</table>

<table>
<tr><td rowspan="8">3주차</td><td>일시</td><td colspan="2">12월 1일 14 : 50 상상파크플러스</td></tr>
<tr><td rowspan="6">참여 학생</td><td>클럽원 정보</td><td>참석 여부</td></tr>
<tr><td>김하경(2394010)</td><td>O</td></tr>
<tr><td>윤서현(2394048)</td><td>O</td></tr>
<tr><td>정윤녕(2394043)</td><td>O</td></tr>
<tr><td>조수빈(2394025)</td><td>O</td></tr>
<tr><td>형유림(2394016)</td><td>O</td></tr>
<tr><td>진도</td><td>도서명: 위험한 인공지능</td><td>진도페이지: p.113 ~ p.171</td></tr>
<tr><td>토론 내용</td><td colspan="2"> 
[좌측부터 김하경, 정윤녕, 형유림, 윤서현, 조수빈]

김하경: 4장 자연어 처리에 대해 발표하며 그중 모델 보안을 중심으로 발표했습니다. 프롬프트 인젝션 등 언어 모델이 악용될 수 있는 위협을 소개하며, 사용자 요구를 악의적으로 조작하는 공격에 대비해 입력 검증과 정책 설정이 중요하다고 설명했습니다.

윤서현: 언어 모델이 실제 대화에서 거짓 정보를 생성하거나 민감 데이터 노출을 유발할 수 있다며, 안전 장치 부재 시 피해가 커질 수 있는 점을 토론에서 강조하며 의견을 제시했습니다.

정윤녕: 역할 기반 지시를 악용해 모델을 속이는 공격 사례가 인상 깊었다고 언급했습니다. 단순한 차단보다 공격 패턴을 학습하는 방식이 필요하며, 사용자 경각심도 함께 강화되어야 한다고 말했습니다.

조수빈: 언어 모델이 범죄 수행 방법 제시나 악성 코드 작성에 악용될 수 있다는 점을 우려했습니다. 정보 제공 제한이나 출력을 제어하는 기술이 강화되어야 한다고 의견을 밝혔습니다.

형유림: 대규모 데이터로 학습해 정확도를 높인 모델일수록 공격 표면이 넓어진다고 지적했습니다. 안전한 데이터 관리와 요청 필터링 기술이 함께 발전해야 한다고 설명하며 토론에 참여했습니다.</td></tr>
</table>

<table>
<tr><td rowspan="8">4주차</td><td>일시</td><td colspan="2">12월 2일 19 : 00 코딩라운지 103호</td></tr>
<tr><td rowspan="6">참여 학생</td><td>클럽원 정보</td><td>참석 여부</td></tr>
<tr><td>김하경(2394010)</td><td>O</td></tr>
<tr><td>윤서현(2394048)</td><td>O</td></tr>
<tr><td>정윤녕(2394043)</td><td>O</td></tr>
<tr><td>조수빈(2394025)</td><td>O</td></tr>
<tr><td>형유림(2394016)</td><td>O</td></tr>
<tr><td>진도</td><td>도서명: 위험한 인공지능</td><td>진도페이지: p.2 ~ p.28</td></tr>
<tr><td>토론 내용</td><td colspan="2"> 

[좌측부터 윤서현, 정윤녕, 조수빈, 석병진 교수님, 형유림, 김하경]

김하경: 멀티모달 AI가 다양한 정보를 결합해 판단할 수 있어 편의성이 높아지지만, 잘못된 결합으로 오해를 일으킬 가능성도 있다고 언급하며 기술 신뢰성 확보가 중요하다고 말했습니다.

윤서현: 1장 격변하는 AI 시장 중 멀티모달 AI와 공간지능을 중심으로 발표했습니다. 텍스트·영상·음성 정보를 함께 처리하여 실제 환경을 더 잘 이해할 수 있지만, 감시 기술로 악용될 경우 사생활 침해가 우려된다고 설명했습니다.

정윤녕: 사용자 위치나 행동을 AI가 인지하는 기술이 보안 개선에는 도움이 되지만 과도한 추적이 될 수 있어 데이터 관리 기준이 명확해야 한다고 의견을 제시하며 토론에 참여했습니다.

조수빈: 실세계 정보를 기반으로 판단하는 시스템이 오작동할 경우 사고로 이어질 수 있다고 우려했습니다. 특히 자율주행 등에서 안전을 확보하기 위한 검증 절차가 중요하다고 강조했습니다.

형유림: 멈춤선이나 장애물을 잘못 인식하는 사례를 보며 공간지능의 한계를 지적했습니다. 고도화된 센서 융합이 필요하며, 시스템의 책임 구조를 명확히 해야 한다고 의견을 남겼습니다.</td></tr>
</table>

	No.	클럽원 정보	후기 내용
활동 후기	1	김하경 (2394010)	이번 독서클럽 활동을 통해 인공지능 기술이 우리의 일상과 보안에 얼마나 깊이 연결되어 있는지 알게 되었습니다. 팀원들의 발표를 통해 다양한 시각을 접할 수 있었고, 교수님께 받은 피드백을 바탕으로 기술을 더 깊이 이해할 수 있었습니다. 특히 토론 과정에서 단순히 내용을 학습하는 것을 넘어, 위험 요소와 대응 방안을 함께 고민할 수 있었던 점이 큰 배움이 되었습니다. 앞으로도 인공지능 보안 분야에 대해 지속적으로 탐구하고 싶습니다.
	2	윤서현 (2394048)	AI 관련 서적을 바탕으로 내용을 조사하고 정리하면서 기술의 흐름을 더욱 명확히 이해할 수 있었습니다. 이후 이를 교수님과 팀원들과 공유하며 서로의 의견을 나누는 과정은 제 학습에 큰 도움이 되었을 뿐만 아니라 협업의 중요성을 다시 한 번 느끼게 해준 값진 경험이었습니다.
	3	정윤녕 (2394043)	이번 독서토론에서는 읽은 내용을 바탕으로 직접 발표해보는 시간을 가졌습니다. 특히 책의 주제인 AI에 대해 자료를 조사하고 준비하면서 관련 지식을 깊이 있게 탐구할 수 있는 좋은 기회였습니다. 발표 후에는 교수님의 구체적인 첨언 덕분에 전공 지식을 한층 더 확장할 수 있어 매우 유익했습니다.
	4	조수빈 (2394025)	이번 독서클럽은 '위험한 인공지능' 책을 읽은 후 각자 원하는 주제를 발표하는 방식으로 진행하였다. 이를 통해 수업에서는 자세히 배우지 않았던 인공지능이 어떻게 기능하게 되는지, 어떤 부분이 위험한지 학우들의 발표를 통해 더 쉽게 이해할 수 있었다. 또한 지도교수님과 함께 진행하면서 교수님께서 설명해주시는 피드백도 학습하는데 매우 큰 도움이 되었다.
	5	형유림 (2394016)	인공지능과 보안이라는 두 분야를 합친 'AI 보안'은 최근 큰 관심을 받고 있다. AI 기술이 매우 빠르게 발전하고, 인간에게 가장 큰 도우미로 자리잡은 현재 AI는 보안의 가장 큰 위협 중 하나로 거론된다. 보안학과 전공자로서 안 들여다볼 수 없는 분야다. 『위험한 인공지능』은 AI 보안을 공부하기 위한 입문서로 딱 적절한 책이었다. 본 책은 AI로 보안을 위협하는 기술과, 이를 방어하는 기술들을 폭넓고 자세하게 기술한다. 친구들과 교수님과 읽고 토론하면서 새로 알게 된 것들이 매우 많았다. 미처 다 읽지 못한 부분까지 완독하고 싶다.