기본연구 | 25-08

생성형 AI의 생산성 분석

민대홍

2025. 12

KISDI 정보통신정책연구원
KOREA INFORMATION SOCIETY DEVELOPMENT INSTITUTE

기본연구 | 25-08

생성형 AI의 생산성 분석

민대홍

2025. 12

KISDI 정보통신정책연구원
KOREA INFORMATION SOCIETY DEVELOPMENT INSTITUTE

서 언

생성형 AI 서비스의 등장과 확산에 따라 세계 주요국 및 우리나라는 생성형 AI를 새로운 성장동력으로 인식하고 경제 전반에서 생성형 AI의 효과적인 도입을 위한 다양한 정책적 노력을 기울이고 있습니다. 생성형 AI의 경제적 파급효과를 추산하는 연구들이 글로벌 컨설팅 회사와 학계를 중심으로 이루어졌으며, 다양한 업무를 대상으로 진행된 실험연구들을 통해 생성형 AI의 생산성 향상 효과에 대해 전반적으로 합의가 이루어지고 있는 상황입니다. 하지만, 생성형 AI가 어떠한 경로를 통해서 생산성을 증가시키는지에 대한 연구는 아직 활발하게 진행되고 있지 않습니다. 미래에 성장을 견인해 나갈 생성형 AI의 잠재력을 극대화하기 위해서는 생성형 AI의 생산성 향상 경로에 대한 이해가 선행되어야 합니다. 이에 본 보고서는 생성형 AI의 잠재적인 생산성 향상 경로를 검증하기 위한 실험을 고안 및 진행하고 수집된 실험데이터를 분석하여, 향후 생성형 AI 효율적인 활용을 위한 정책 수립에 필요한 기초자료를 제공하고자 하였습니다.

본 연구는 정보통신정책연구원의 민대홍 연구위원이 총괄책임을 맡아 수행하였습니다. 본 연구를 시작으로 생성형 AI의 생산성 향상 경로에 관한 논의가 폭넓고 활발하게 진행될 수 있기를 기대합니다. 아울러 본 보고서의 분석 내용이 급속히 변화하는 AI 전환기에 보다 실효성있는 생성형 AI 관련 정책 방향을 설정하는데 기여할 수 있기를 기원합니다.

2025년 12월

정보통신정책연구원

원 장 이 상 규

목 차

서 언 ····· 1

요약문 ····· 9

제1장 서 론 ····· 11

1. 연구의 배경 및 필요성 ····· 11
2. 연구의 목표 ····· 12

제2장 생성형 AI 관련 선행 실험연구 분석 ····· 14

제3장 생성형 AI의 생산성 분석 ····· 46

제1절 실험설계 ····· 46

1. 실험 설계 전 고려 사항 ····· 46
2. 실험 설계 ····· 47
3. 검증 가능 가설 ····· 52

제2절 실험 결과 분석 ····· 57

1. 실험 실사 ····· 57
2. 과업 결과물의 평가 ····· 61
3. 실험 분석 결과 ····· 63
4. 소결 ····· 84

제4장 결 론 ····· 87

참고문헌 ····· 89

부 록 ····· 92

Abstract ····· 110

표 목 차

〈표 2-1〉 선행 연구 요약 ····· 43
〈표 3-1〉 실험설계 요약 ····· 52
〈표 3-2〉 응답자 일반 특성과 AI 활용 특성 ····· 63
〈표 3-3〉 과업1의 품질: 대조군(hteam)과 실험군2(hindv) ····· 65
〈표 3-4〉 과업2의 품질: 대조군(hteam)과 실험군2(hindv) ····· 66
〈표 3-5〉 과업1의 품질: 대조군(hteam)과 실험군3(aiindv) ····· 68
〈표 3-6〉 과업1의 소요시간: 대조군(hteam)과 실험군3(aiindv) ····· 69
〈표 3-7〉 과업2의 품질: 대조군(hteam)과 실험군3(aiindv) ····· 70
〈표 3-8〉 과업2의 소요시간: 대조군(hteam)과 실험군3(aiindv) ····· 70
〈표 3-9〉 과업1의 품질: 실험군2(hindv)와 실험군3(aiindv) ····· 72
〈표 3-10〉 과업1의 소요시간: 실험군2(hindv)와 실험군3(aiindv) ····· 73
〈표 3-11〉 과업2의 품질: 실험군2(hindv)와 실험군3(aiindv) ····· 74
〈표 3-12〉 과업2의 소요 시간: 실험군2(hindv)와 실험군3(aiindv) ····· 74
〈표 3-13〉 과업1의 품질: 대조군(hteam)과 실험군1(aiteam) ····· 76
〈표 3-14〉 과업1의 소요 시간: 대조군(hteam)과 실험군1(aiteam) ····· 77
〈표 3-15〉 과업2의 품질: 대조군(hteam)과 실험군1(aiteam) ····· 77
〈표 3-16〉 과업2의 소요 시간: 대조군(hteam)과 실험군1(aiteam) ····· 78
〈표 3-17〉 과업1의 품질: 실험군1(aiteam1)과 실험군3(aiindv) ····· 81
〈표 3-18〉 과업1의 소요 시간: 실험군1(aiteam1)과 실험군3(aiindv) ····· 81
〈표 3-19〉 과업2의 품질: 실험군1(aiteam1)과 실험군3(aiindv) ····· 82
〈표 3-20〉 과업2의 소요 시간: 실험군1(aiteam1)과 실험군3(aiindv) ····· 83
〈표 3-21〉 결과 요약 ····· 84

그 림 목 차

[그림 2-1] 과업 지시문 ····· 15
[그림 2-2] 시간당 상담 건수 ····· 21
[그림 2-3] 상담 건당 소요시간 분포 ····· 21
[그림 2-4] 상담 결과 고객 평가점수 ····· 22
[그림 2-5] 코딩 과업 지시서 ····· 23
[그림 2-6] 실험군에게 제시된 GPT-4에 대한 소개 ····· 25
[그림 2-7] 실험 참가 그룹 간 참신성, 효용성 지수 비교 ····· 29
[그림 2-8] 1 GenAI idea 그룹과 5 GenAI idea 그룹 간의 참신성, 효용성 비교 ····· 29
[그림 2-9] 그룹 간 이야기의 유사성 정도 분포 ····· 30
[그림 2-10] 이메일 과업에 주어진 이메일 내용 ····· 32
[그림 2-11] 장학생 그룹과 비장학생 그룹의 생산성 향상 효과 비교 ····· 37
[그림 2-12] 실험 결과 ····· 39
[그림 2-13] 평가점수 분포 비교 ····· 41
[그림 2-14] 과업 평가결과 비교 ····· 42
[그림 3-1] 실험에서 사용된 설명문 ····· 50
[그림 3-2] 실험 화면 스크린샷: 생성형 AI 활용 ····· 59
[그림 3-3] 실험 화면 스크린샷: 생성형 AI 비활용 ····· 60
[그림 3-4] 과업1 결과물의 품질 평가를 위한 채점기준 ····· 61
[그림 3-5] 과업2 결과물의 품질 평가를 위한 채점기준 ····· 62
[그림 3-6] 과업1 결과물의 품질: 대조군(hteam) vs. 실험군2(hindv) ····· 65
[그림 3-7] 과업2 결과물의 품질: 대조군(hteam) vs. 실험군2(hindv) ····· 66

[그림 3-8] 과업1 결과물의 품질: 대조군(hteam) vs. 실험군3(aiindv) ····· 69
[그림 3-9] 과업2 결과물의 품질: 대조군(hteam) vs. 실험군3(aiindv) ····· 71
[그림 3-10] 과업1 결과물의 품질: 실험군2(hindv) vs. 실험군3(aiindv) ···· 73
[그림 3-11] 과업2 결과물의 품질: 실험군2(hindv) vs. 실험군3(aiindv) ···· 74
[그림 3-12] 과업1 결과물의 품질: 대조군(hteam) vs. 실험군1(aiteam) ··· 76
[그림 3-13] 과업2 결과물의 품질: 대조군(hteam) vs. 실험군1(aiteam) ··· 78
[그림 3-14] 과업1 결과물의 품질: 실험군1(aiteam) vs. 실험군3(aiindv) · 82
[그림 3-15] 과업2 결과물의 품질: 실험군1(aiteam) vs. 실험군3(aiindv) · 83

요 약 문

1. 제 목

생성형 AI의 생산성 분석

2. 연구 목적 및 필요성

생성형 AI의 활용이 생산성 증가로 이어진다는 결과는 최근에 수행된 다수의 연구에서 확인되었다. 하지만 이러한 생산성 향상이 어떠한 경로를 통해 이루어지는지에 대한 논의는 활발하지 않다. 이에 본 연구는 생성형 AI의 생산성 향상 경로 중 분업과 관련된 경로에 주목하여, 생성형 AI의 활용과 분업의 상호작용을 실험을 통해 분석하고자 한다.

3. 주요 연구내용 및 정책적 시사점

실험은 생성형 AI의 활용 여부와 분업 유무가 상호작용할 수 있도록 2×2 설계로 구성되었다. 생성형 AI 활용 여부와 분업의 유무에 따라 네 집단이 가상의 제품에 관한 마케팅 과업을 수행하였다. 마케팅 과업은 2개 과업으로 구성되고 순차적으로 수행되는데, 세부 과업1과 세부 과업2는 분업을 통해 수행될 수도, 분업 없이 한 명의 개인이 연속적으로 수행할 수 있도록 설계하였다. 만 19세 이상 만 60세 미만 대한민국 성인 대상 온라인 실험이 실시되어 총 553개의 유효응답이 수집되었다.

실험 결과, 먼저 생성형 AI가 활용되지 않는 상황에서, 분업을 통해 과업을 실

시한 집단의 과업 결과물의 품질이 그렇지 않은 집단보다 낮게 나타나 기존 문헌에서 논의되었던 정보손실 및 정보해석 비용으로 인한 분업의 비효율성을 재확인하였다.

생성형 AI의 활용이 분업을 소멸시키는 상황에서 생산성 변화를 측정하기 위하여, 생성형 AI 없이 분업을 통해 과업을 수행한 집단과 생성형 AI를 활용하면서 분업 없이 과업을 수행한 집단을 비교한 결과, 생성형 AI의 활용, 분업의 소멸, 그리고 생산성 향상으로 이어지는 경로를 확인하였다. 즉, 생성형 AI로 인해 종전에는 분업으로만 수행할 수 있던 과업이 분업 없이 수행할 수 있게 될 경우, 분업이 생산성에 미치는 부정적인 효과가 제거되고 생성형 AI 활용 그 자체로 인한 생산성 향상으로 인해 생산성 증가가 가능함을 확인하였다.

마지막으로 생성형 AI가 분업에 미치는 영향을 살펴보기 위해, 생성형 AI가 활용되는 상황에서 분업으로 과업을 수행한 집단과 그렇지 않은 집단을 비교한 결과, 두 집단 간의 과업 결과물 품질에서 유의미한 차이는 없는 것으로 나타났다. 이는 분업이 존재하는 환경에서 생성형 AI가 활용될 경우, 앞서 확인하였던 분업의 비효율성을 상쇄하는 것으로 이해할 수 있다. 해당 결과는 순차적 분업 상황에서 단순히 AI가 노동을 대체하여 증가하는 생산성 증분과 AI가 노동을 대체하지 않고 보완하여 증가하는 생산성 증분이 크게 다르지 않을 수 있음을 시사한다.

제 1 장 서 론

1. 연구의 배경 및 필요성

생성형 AI는 등장과 동시에 다양한 산업에 빠르게 도입·활용되며, 새로운 성장 동력을 제공할 기술로 주목받고 있다. Mckinsey(2023)에 따르면, 생성형 AI의 경제적 파급효과는 최소 2.6조 달러에서 최대 4.4조 달러에 이를 것으로 추정된다. 과학기술정보통신부와 Bain&Company의 공동연구에서는 우리나라를 중심으로 생성형 AI의 경제적 효과를 분석했으며, 그 결과 매출 증대와 비용 절감 효과 등을 포함한 생산 역량의 잠재력이 약 310조 원에 달하는 것으로 나타났다.[1] 이러한 분석은 생성형 AI가 기존 산업을 혁신하고 비용을 절감함으로써 우리나라가 직면한 저출산, 고령화, 저성장 문제를 극복할 수 있는 핵심 기술임을 시사한다.

이와 같은 배경에서, 우리나라는 해외 주요국들의 디지털 기술 패권 경쟁 속에서 글로벌 경쟁력을 확보하기 위해 다양한 전략을 수립해 왔다. 2022년에는 '디지털 전략'을, 2024년에는 'AI·디지털 혁신성장 전략'을 발표하며 국가 역량을 집중하고 있다. 미국은 〈반도체와 과학법(2022년 8월)〉을 통해 AI를 포함한 10대 핵심기술 분야에 총 2,800억 달러 규모의 투자를 단행했고, 중국은 2021년 '디지털 중국'을 선언하며 인공지능을 포함한 7대 중점 산업을 육성함으로써 신성장 동력 확보에 경주하고 있다. 유럽연합(EU)은 디지털 주도권 확보와 혁신을 위해 '2030 디지털 나침반(Digital Compass)'을 수립하고 기업들의 인공지능 활용을 촉진하고 있으며, 영국은 2022년 '영국 디지털 전략'을 발표해 AI 고급 인재 양성과 AI 기반의 공공 혁신에 속도를 내고 있다. 이에 발맞추어 우리나라도 관계부처

1) 'AI 일상화'에 올해 7100억 원 투입…AI로 건강관리·돌봄 등 지원. (2024년 4월 4일). 대한민국 정책브리핑.
https://www.korea.kr/news/policyNewsView.do?newsId=148927861

합동으로 2022년 '대한민국 디지털 전략'을 수립해 디지털 경쟁력 강화를 위한 정책을 추진해 왔으며, 2024년에는 후속 전략으로 'AI·디지털 혁신성장 전략'을 통해 AI 확산에 7,000억 원을 투자하고, AI G3로 도약하기 위한 국가적 역량을 결집하고 있다.

하지만 지금까지의 생성형 AI 관련 생산성 향상 연구들은 방법론과 데이터의 투명성이 부족한 실정이다. 글로벌 컨설팅 그룹들(Mckinsey, KPMG, Bain& Company 등)이 수행한 연구에서는 다양한 활용 사례를 기반으로 생산성 향상 정도를 추정하고 있지만, 그 구체적인 방법론이나 데이터는 공개되지 않고 있다. 신기술의 경제적 파급 효과에 대한 대안적 추정 방식이 부재한 상황에서, 추정 방법에 대한 명확한 이해 없이도 정책 수립에 활용할 수밖에 없는 현실이다.

이러한 한계 속에서 생성형 AI의 중요성이 높아지는 현시점에서, 투명하고 설명 가능한 방법론에 기반한 데이터와 추정치의 확보가 필요하다. 이에 실험경제학적 방법론을 활용하여, 실험 설계 단계부터 데이터 수집에 이르기까지 전 과정을 투명하게 진행함으로써, 생성형 AI의 생산성 향상 추정치에 대한 신뢰도를 높이고자 한다. 또한 최근 다양한 연구에서는 실험경제학적 방법론이 활용되고 있으나, 대부분 생성형 AI의 생산성 향상을 확인하였을 뿐, 어떠한 경로로 인해 생산성이 향상되는지에 대한 연구는 찾아보기 힘들다. 이에 본 연구에서는 생성형 AI가 생산성 향상을 이끌어내는 경로 중 하나로 분업과 생성형 AI 활용의 상호작용에 집중하여 생성형 AI의 생산성 향상 경로에 집중한 실험 연구를 진행하고자 한다.

2. 연구의 목표

이에 본 연구는 생성형 AI의 생산성과 관련된 문헌연구를 통해 해당 기술의 다양한 활용 사례와 생산성과 관련된 자료를 조사하고, 이를 바탕으로 실험경제학적 방법론을 활용하여 생성형 AI의 생산성 증가 경로 파악 및 증가분을 추정하는 데 목적이 있다. 이를 통해 생성형 AI에 대한 증거 기반 정책 입안의 기초자료를 마련하고자 한다.

우선, 생성형 AI의 생산성 향상 효과 및 경제적 효과와 관련된 기존 연구들을 종합적으로 검토한 후, 기존 연구들이 갖는 한계를 보완할 수 있는 실험 설계를 통해 새로운 데이터를 수집하고자 한다. 앞서 언급하였듯이 기존 문헌에서 다루지 않았던 생성형 AI의 생산성 향상 경로를 분석함으로써 달성하고자 한다. 특히 많은 생산이 분업으로 이루어지는 현대 경제에서 생성형 AI 활용에 따라 노동이 대체되는 것은 곧 분업의 소멸 또는 인간의 결과물이 아닌 생성형 AI와의 분업을 의미하게 된다. 본 연구에서는 분업과 생성형 AI의 상호작용을 살펴보기 위한 실험을 설계하고, 이를 통해 생성형 AI가 분업과 상호작용하며 어떠한 경로로 생산성을 변화시키는지 그 경로를 살펴보고자 한다

제 2 장 생성형 AI 관련 선행 실험연구 분석

본 장에서는 실험연구를 활용하여 생성형 AI의 생산성을 추정한 연구들을 살펴보고, 이를 분석함으로써 추후 실험설계에 활용하고자 한다.[2)]

가. Noy and Zhang(2023)

Noy and Zhang(2023)은 실험연구를 진행하여 생성형 AI의 생산성 향상을 과업완료 속도와 과업 결과물의 질을 활용하여 추정하였다. 총 453명의 대학졸업자를 대상으로 중급 글쓰기(midlevel writing skill)와 관련된 과업을 수행하도록 하는 한편, 이들 중 절반의 인원을 무작위로 선택하여 과업 수행에서 ChatGPT를 활용하도록 하였다. 그 결과, ChatGPT를 활용한 집단이 그렇지 않은 집단에 비하여, 40% 정도 빠르게 과업을 완료하였고 과업 결과물 역시 그렇지 않은 집단에 비하여 18% 정도 향상하는 것으로 나타났다. 또한 ChatGPT를 활용한 집단은 실험 종료 이후에도 생성형 AI를 자신의 업무에 활용하는 인원이 그렇지 않은 집단보다 2주 후까지는 2배, 2달 후에는 약 1.6배 더 많은 것으로 나타났다.

Noy and Zhang(2023)의 주요 연구 목표는 생성형 AI의 생산성 향상 증분, 생성형 AI의 활용이 숙련 노동자와 비숙련 노동자의 불평등도에 미치는 영향, 실험 참가자들이 생성형 AI에 반응하는 양식의 분석이다.

이러한 연구 목표를 달성하기 위해 Noy and Zhang(2023)은 다음과 같은 실험을 설계하고 실행하였다. 실험은 온라인 설문 플랫폼인 프롤리픽(Prolific)에서 실시되었다. 총 453명의 실험 참가자를 모집하였는데 이들은 인사전문가, 컨설턴트, 데이터 분석가, 연구제안 전문가(grant writer), 관리자 직급, 마케터 등 다양

2) 본 절은 민대홍(2024)의 논의와 자료를 바탕으로 일부 내용을 재구성·보완하여 작성하였다.

한 분야의 전문가로 구성되었다. 참가자들은 실험에서 자신의 실제 직무와 가장 가까운 문서작성 과업을 수행하였다. 예를 들어, 마케터들은 특정 제품에 대한 보도자료 작성, 데이터 분석가들은 데이터 분석 계획서(실제 코딩 제외), 연구제안 전문가들은 연구제안서의 커버레터(cover letter), 인사전문가는 민감한 주제(재택 근무 중 헤드폰 착용 의무 등)에 관련한 공지를 전사에 전달하는 이메일 작성, 컨설턴트는 짧은 보고서를 작성하였다. 각 과업은 약 20~30분 정도가 소요되도록 고안되었다. 참가자들이 높은 품질의 문서를 작성하도록 유도하기 위하여, 각 참가자에게 기본금으로 10달러를 지급하고, 과업 결과물의 질에 따라 최고 14달러를 추가로 지급받을 수 있도록 설계하였다. 참가자들이 작성한 과업 결과물의 품질은 동종업종에서 근무 경험이 충분한 전문가들이 평가한 후 평가 결과에 따라 보너스를 지급하였다.

[그림 2-1] 과업 지시문

"Imagine you are a senior manager working for a company, WorkCo, in 2025. Most of the company's employees work from home.

In an attempt to recreate the collaborative office environment while allowing employees to work from home, the company has recently paid for the creation of a virtual office space and sent each employee a virtual-reality headset that allows them to connect to the space. ... However, since receiving the virtual-reality headsets, employees have not used them very much. .... because they prefer the privacy of working alone at home. ... Write a draft of a company-wide email that gently but firmly encourages employees to begin using their head sets."

참가자의 절반이 무작위로 선택되어 과업 수행에서 ChatGPT를 활용하고, 나머지 절반은 생성형 AI를 사용하지 않았다. 두 그룹 간 비교뿐만 아니라, ChatGPT를 활용하는 그룹 내 참가자들의 생산성 변화를 측정하기 위하여, 모든 참가들은 두 개의 과업을 진행하였다. 즉, ChatGPT를 활용한 그룹은, ChatGPT 활용 이

전에 하나의 과업을 그리고 ChatGPT를 활용하여 또 다른 과업을 수행하였고, ChatGPT를 사용하지 않는 그룹은 하나의 과업을 수행한 후, 생산성과는 관련없는 소프트웨어(Overleaf)를 제공받은 후 두 번째 과업을 진행하였다.

생산성 측정을 위해 사용한 척도는 과업 완료 시간과 결과물의 질로, ChatGPT를 활용한 그룹이 그렇지 않은 그룹보다 평균적으로 16분 빨리 과업을 완료하였다. 또한 과업 결과물의 질 역시 ChatGPT를 활용한 그룹이 그렇지 않은 그룹보다 약 0.45점 더 높은 평가 점수를 획득하였다. ChatGPT를 사용한 그룹 내 참가자의 ChatGPT 사용에 따른 생산성 변화는 첫 번째 과업과 두 번째 과업을 비교하여 측정하였는데, 평균적으로 ChatGPT를 사용하지 않은 첫 번째 과업보다 ChatGPT를 사용한 두 번째 과업의 수행시간이 약 10분 정도 적게 소요되었으며, 과업 결과물 역시 약 1~2점 더 높은 평가 점수를 획득하는 것으로 나타났다.

ChatGPT가 생산성 불평등을 완화하는 효과 역시 발견하였는데, ChatGPT를 사용하지 않은 그룹에서는 첫 번째 과업에서 높은 점수를 획득한 참가자가 두 번째 과업에서도 높은 점수를 획득하는 경향이 강하게 나타난 반면, ChatGPT를 활용한 그룹에서는 이러한 경향이 약화되어 나타났다. 즉, ChatGPT를 사용하지 않은 첫 번째 과업에서 낮은 평가를 받은 참가자가 ChatGPT를 활용한 두 번째 과업에서 비교적 높은 점수를 획득할 확률이 증가하였다.

나. Choi and Schwarcs(2023)

Choi and Schwarcs(2023)는 법무(legal works)와 관련된 생성형 AI 활용에 초점을 맞추어 실험 분석을 진행하였다. 생성형 AI가 법학전문대학원 입학시험, 변호사 시험, 그리고 다른 종류의 법무 관련 분석에 탁월한 성능을 보인다는 점에 주목하여, 미네소타대학교(University of Minnesota)의 법학전문대학원에 재학중인 학생들을 대상으로 실험을 진행하였다. 실험에 참가한 학생들은 자신들이 수강 중인 수업의 실제 기말고사 후, GPT-4의 활용법에 대한 교육을 받고 가상의

시험에서 GPT-4를 활용하여 답안을 작성하였다. 실험 결과, 학생들이 GPT-4를 활용하면 객관식 문제의 성적이 그렇지 않은 경우보다 약 29% 정도 상승하였지만, 서술형 문제에서는 큰 차이를 보이지 않았다. 또한, 실제 수업에서 낮은 성적을 기록한 학생은 GPT-4를 활용한 가상의 시험에서 약 40%의 성적 향상을 보인 반면, 첫 번째 시험에서 좋은 성적을 기록한 학생은 GPT-4를 활용한 시험 성적에서 약 20% 하락한 것으로 나타났다.

이뿐만 아니라, Choi and Schwarcs(2023)는 생성형 AI, 인간, 그리고 생성형 AI를 활용하는 인간의 생산성을 비교하였는데, 먼저 학부 신입생들이 수강하는 기본적인 법학 입문 수업의 시험에서 AI는 B+에서 A의 성적을 얻을 수 있었다. 또한, 고급 법학 개념을 배우는 수업의 시험에서는 B에서 A-정도의 성적을 기록하여,[3] 과업과 관련된 정보를 적절히 교육시킨 GPT-4가 일반 학생들이나 GPT-4를 활용한 학생들보다 평균적으로 높은 성적을 얻는 것으로 나타났다.

실험에는 미네소타대학교의 법학전문대학원 재학생들 가운데 보험법을 수강하는 52명의 학생 중 16명, 그리고 미국법 입문 강의를 수강하는 128명의 학생 중 32명이 참가하였다. 실험에는 2022년에 실제로 사용한 시험문제의 일부를 활용하였으며, 이를 완료하는 데는 약 2주의 시간이 주어졌다. 같은 시험문제를 GPT-4 스스로 풀게 한 후, 앞서 참가자들의 시험지와 2022년에 생성형 AI 없이 시험에 응시한 학생들의 시험지를 모두 섞은 뒤에 채점을 진행하여 편향 가능성을 낮추었다.

실험 결과는 각 수업별로 나누어 살펴보았는데, 먼저 객관식 문제와 서술형 문제를 모두 가지고 있는 미국법 입문의 경우, 객관식 문항에서 GPT-4를 활용하여 2023년에 2022년 시험에 응시한 참가자들의 성적이, 생성형 AI의 도움 없이 2022년에 2022년 시험에 실제로 응시한 학생들의 성적의 약 88 백분위인 것으로 나타나, 객관식 문제에서 생성형 AI가 정답률을 유의미하게 올리는 것으로 보

3) 두 시험 모두 GPT-4에 각 강의의 강의노트를 학습시킨 후 시험을 치르게 하였다.

인다. 하지만 미국법 입문의 서술형 문제 또는 서술형 문제로만 구성된 보험법의 시험에서는 유의미한 차이가 발견되지 않았다.

마지막으로 실제 수업에서의 높은 성적을 받은 학생들의 경우 GPT-4를 활용한 가상의 시험에서 이들의 실제 수업에서의 백분위보다 약 20% 감소한 백분위를, 실제로 낮은 성적을 받은 학생들은 GPT-4를 활용한 가상의 시험에서 실제 수업에서의 백분위보다 약 40% 증가한 백분위를 기록하여 다른 문헌과 비슷한 결과를 도출하였다.

흥미롭게도 이 연구에서는 GPT-4가 스스로 가상의 시험에 응시한 답안 역시 평가하고 이를 GPT-4의 활용 없이 응시한 학생의 성적, 그리고GPT-4를 활용한 학생의 성적과 비교하였다. 미국법 입문의 객관식 문제와 서술형 그리고 보험법의 서술형 시험 모두에서 GPT-4가 GPT-4 없이 시험을 치른 학생들의 성적(2022년 실제 수업에서 학생들이 응시한 성적)보다 높은 성적을 받았다. 서술형 부분에서의 시험성적은 생성형 AI 활용 전략(prompt strategy)에 따라 다른 결과를 보였는데, 상대적으로 간단한 미국법 입문의 서술형 시험에서는 활용 전략에 상관없이 GPT-4의 시험 성적이 높았지만, 복잡한 사고가 요구되는 보험법에서는 간단한 활용 전략의 경우 다른 전략들에 비하여 더 낮은 점수를 도출하였다.

Choi and Schwarcs(2023)는 기존에 시도하지 않았던 생성형 AI의 성능과 생성형 AI를 활용하는 인간의 능률을 비교하였다는 점에서 인상적이나, 실험 참가자의 수가 매우 적고, 이들이 실험에 참가하여 최선을 다할 이유가 없다는 점에서 한계를 가진다. 하지만, 참가한 모든 학생들이 GPT-4의 활용 방법에 대한 교육을 수료한 후에 실험을 진행하였다는 점에서는 GPT-4 활용 능력에 따른 생산성 증분의 편차를 없애 데이터의 질을 높였다고 할 수 있다.

다. Brynjolfsson et al.(2023)

Brynjolfsson et al.(2023)은 실제 고객상담센터에서 근무 중인 5,172명을 대

상으로 생성형 AI가 이들의 직무 생산성 증분을 추정할 수 있는 실험을 진행하였다. 실험 결과 생성형 AI를 활용할 경우 이들이 시간당 처리하는 상담 건수로 측정한 직무 생산성이 평균 약 15% 증가한 것으로 나타났다. 이러한 생산성 증분 효과는 근무자들 간에 상당히 상이하였는데, 비숙련 근무자들의 경우 상담 해결 속도와 상담의 품질이 모두 증가하였으나, 숙련 근무자들의 상담 해결 속도는 미약하게 증가하는 반면 상담의 품질은 소폭 하락하는 것으로 나타났다. 또한, 생성형 AI 활용은 영어가 모국어가 아닌 근무자들의 영어를 더욱 유창하게 하였으며, 생성형 AI 활용이 가장 효과적인 부분은 실제 근무자들은 흔히 경험하지 못하지만 생성형 AI는 데이터를 통해 해답을 알고 있는 문제들로, 극단적이지 않지만 잘 일어나지 않는 문제들에 대한 상담으로 나타났다. 실제 생성형 AI 활용은 근무자들의 직무 경험에 긍정적인 영향을 미치는 것으로 나타났는데, 생성형 AI를 활용함에 따라 응대 고객이 더 예의바르게 반응하거나, 상위 직급자와 직접 이야기하기를 요청하는 건수가 줄어들었다.

Brynjolfsson et al.(2023)은 미국 중소기업을 대상으로 경영 관련 소프트웨어를 제공하는 기업의 고객상담업무 담당자를 대상으로 연구를 진행하였다. 고객상담은 주로 해당 기업이 제공하는 소프트웨어와 관련된 문제를 해결하는 것으로, 평균 약 40분이 소요되며, 주로 온라인 채팅으로 상담이 이루어진다. 이러한 고객상담업무를 수행하기 위해서는 소프트웨어에 대한 이해, 문제해결 능력, 그리고 불편한 감정을 가진 고객을 응대하는 능력이 요구된다. Brynjolfsson et al.(2023)은 소수의 상담업무 담당자들을 대상으로 소규모 RCT를 진행한 뒤, 고객상담 업무 담당자들에게 생성형 AI에 대한 교육을 마친 후 생성형 AI에 대한 접근권한을 주는 방식으로 실험을 진행하였다. 소규모 RCT 이후, 2020년 10월부터 생성형 AI에 대한 접근을 점차 허용하기 시작하여, 2021년 5월에 예산 제약하에서 허용된 생성형 AI 접근 수의 100%를 달성하였다. 따라서 실험 및 데이터 수집 기간은 총 6개월이고, 6개월 동안 고객상담 업무 담당자들의 과업 결과물을 관찰하며, 생성형 AI 활용이 과업 결과물에 미치는 영향을 상담소요시간, 상담결

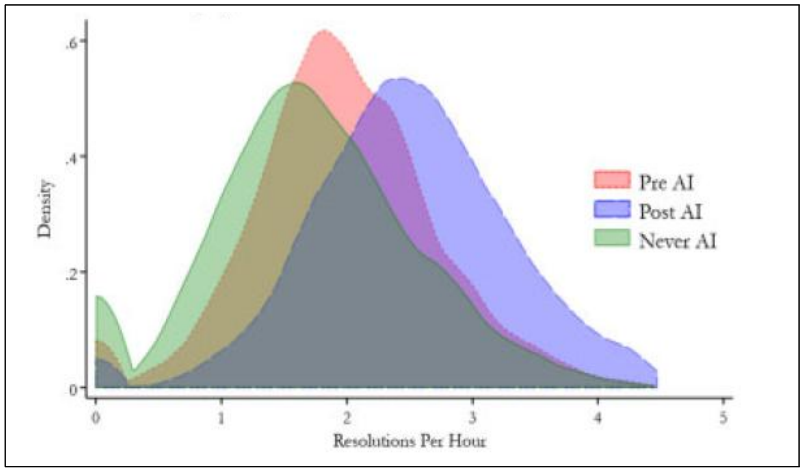
과점수 등을 활용하여 측정하였다. 총 5,172명의 고객상담 담당자들 중 실험이 중단되는 시점까지 생성형 AI를 활용한 담당자는 총 1,340명, 그렇지 않은 담당자는 3,517명이다. 6개월간 순차적으로 담당자들에게 생성형 AI 활용을 허용하였기 때문에, 데이터상으로 총 3개 그룹이 존재하는데, 생성형 AI를 활용하지 않은 3,517명과 실험 종료 전까지 생성형 AI를 활용할 수 있게 된 담당자 중 아직 접근이 허용되지 않은 채로 근무 중인 그룹, 그리고 실제로 생성형 AI를 활용하여 근무하고 있는 그룹이다.

생성형 AI의 도입이 고객상담업무 담당자의 과업 결과물에 미치는 영향을 측정하기 위하여 수집된 데이터를 바탕으로 이중차분법을 활용하였는데, 개인 생산성의 상이성, 계절적 요인을 통제하기 위하여 개인고정효과와 계절효과를 고려한 모형을 사용하였다.

이들 세 그룹의 생산성을 평균적으로 살펴본 결과는 [그림 2-2]와 같이 상당히 명료하다.

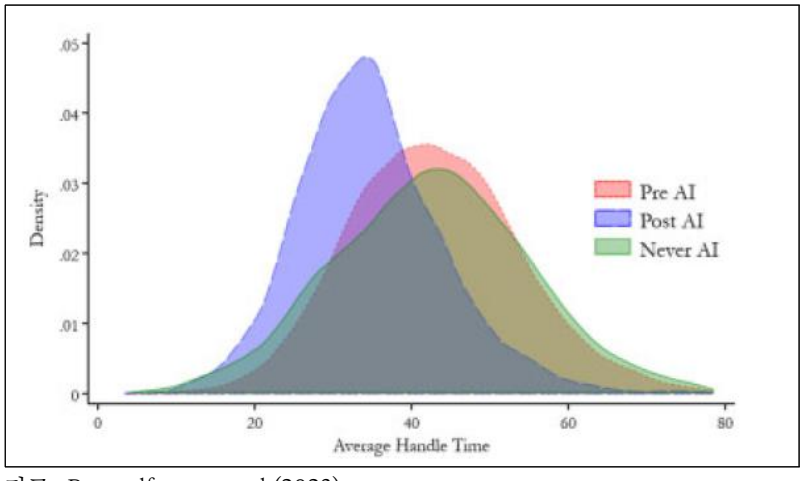
먼저 시간당 완료한 상담 건수를 살펴보면 생성형 AI를 활용한 그룹(Post AI)은 시간당 약 2.5건의 상담을 완료한 반면, 생성형 AI를 활용할 기회가 없던 그룹은 시간당 1.7건의 상담을 완료한 것으로 나타났다. 생성형 AI를 활용하는 그룹이지만 아직 활용이 이루어지지 않은 시점(Pre AI)에 측정한 시간당 완료 상담 건수는 약 2.0건으로, 이는 이미 어느 정도 생산성이 높은 담당자들이 생성형 AI 활용에도 더 적극적이라는 의미로, 생성형 AI 활용에 대한 접근을 획득하기 위해서는 3시간의 교육 이수 등이 부과된다는 점에서 생성형 AI 활용에 지원한 담당자들이 자신의 업무에 대해 더 적극적이고, 따라서 더 높은 생산성을 가지고 있다고 추론할 수 있다.

[그림 2-2] 시간당 상담 건수



자료: Brynjolfsson et al.(2023)

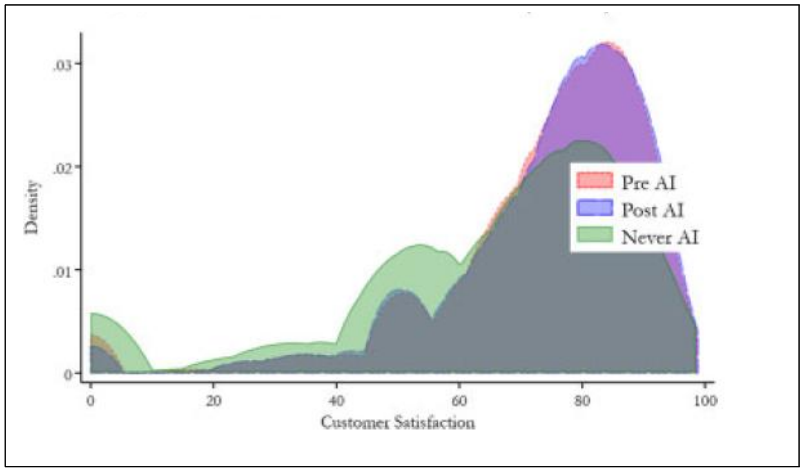
[그림 2-3] 상담 건당 소요시간 분포



자료: Brynjolfsson et al.(2023)

다음으로 상담 건당 소요시간을 살펴보면, 생성형 AI를 활용하지 않은 두 그룹(Never AI와 Pre AI)의 소요시간은 약 40분인 반면, 생성형 AI를 활용한 그룹의 소요시간은 평균 약 35분으로 나타나 더 명확한 생산성 향상을 보여준다.

[그림 2-4] 상담 결과 고객 평가점수



자료: Brynjolfsson et al.(2023)

마지막으로 상담 결과에 대한 고객의 평가점수에서는 생성형 AI를 실험 종료까지 사용하지 않았던 그룹(Never AI)보다 생성형 AI를 실험 종료 전에 사용한 그룹(Pre AI와 Post AI)이 더 높은 점수를 받는 것으로 조사되었으며, 생성형 AI를 실험 종료 전에 사용한 그룹 내 Pre AI 그룹과 Post AI 그룹의 차이는 거의 없는 것으로 나타났다.

라. Peng et al.(2023)

Peng et al.(2023)은 소프트웨어 개발자를 대상으로 생성형 AI의 생산성 증분을 측정하기 위한 실험을 진행하였다. 소프트웨어 개발자들 중 생성형 AI를 활용한 그룹은 그렇지 않은 그룹보다 동일한 과업을 약 55.8% 더 빨리 완료한 것으로 나타났다.

Peng et al.(2023)은 95명의 소프트웨어 개발자들을 모집한 후 무작위로 생성형 AI를 활용할 그룹과 그렇지 않은 그룹으로 나누었다. 소프트웨어 개발자들은 자신의 직무와 관련된 HTTP 서버를 JavaScript로 작성하는 과업을 수행하였다.

참가자들 중 생성형 AI를 활용할 그룹은 GitHub Copilot에 대한 소개 영상 시청 후 해당 생성형 AI에 대한 접근을 획득한 후에 과업을 시작하였다. 두 그룹 모두 자신이 종전 업무에 사용하던 방법, 예를 들어 인터넷 검색 또는 특정 분야에 대한 질문을 해결하는 커뮤니티(Stack Overflow) 등을 사용하는 것에 제약을 두지 않았다. 따라서 두 그룹의 차이는 생성형 AI를 활용할 수 있는지 없는지에만 있다.

총 95명의 참가자 중 45명이 생성형 AI를 활용하는 그룹에, 나머지 50명이 그렇지 않은 그룹에 무작위로 배정되었고, 실험 결과, 생성형 AI를 활용하는 그룹과 그렇지 않은 그룹에서 각각 35명의 참가자가 해당 과업을 완료하였다. 생성형 AI를 활용하는 그룹의 과업 완료시간은 약 71분, 그렇지 않은 그룹은 약 161분이 소요되어, 생성형 AI를 활용하는 그룹이 55.8% 더 빠르게 과업을 완료한 것으로 나타났다.

[그림 2-5] 코딩 과업 지시서

```
/*

Welcome to the GitHub Copilot study!
Your mission, should you choose to accept it:

Write a toy HTTP server engine.
You do not need to actually handle network requests.
Just parse a request string and return a response string.

This is a good reference for the structure of HTTP requests and responses:
https://developer.mozilla.org/en-US/docs/Web/HTTP/Messages

You must parse the request headers, but you do not need to validate them.
You must return a valid HTTP response with a correct content length.
You must serve files from the content/ directory as the root.
  Meaning, `GET /index.html HTTP/1.1` should return content/index.html.
  You do not need to handle `GET /`

Return HTTP status 404 if the file is not found.
Return HTTP status 400 if the request is invalid.
Return HTTP status 405 for any HTTP method other than GET.
Return HTTP status 200 with the contents of the file if the request is valid.

A proper HTTP response with no body has two trailing newlines:
  `HTTP/1.1 404 Not Found\n\n`

A proper HTTP response with a body has a Content-Length header and a body:
  `HTTP/1.1 200 OK\nContent-Length: ${THELENGTH}\n\n${THEFILEBODY}`
*/
```

자료: Peng et al.(2023)

실험 결과와 실험 후 설문을 통해 수집한 사회인구학적 특성을 활용하여 생성형 AI로 인한 생산성 향상 증분의 상이성에 대하여 분석하였는데, 생성형 AI로 인한 생산성 향상 효과는 숙련된 개발자일수록, 코딩 업무가 과중할수록, 그리고 나이가 많은 개발자일수록 높게 나타났다.

이 연구는 참가자들에게 실험 종료 후 설문을 진행하였는데, 생성형 AI를 사용한 그룹과 그렇지 않은 그룹 모두 생성형 AI가 자신의 생산성을 약 35% 증가시킨 또는 증가시킬 것으로 예상한다고 대답하여 실제 생산성 증분인 55.8%보다 더 낮게 평가하고 있는 것으로 나타났다.

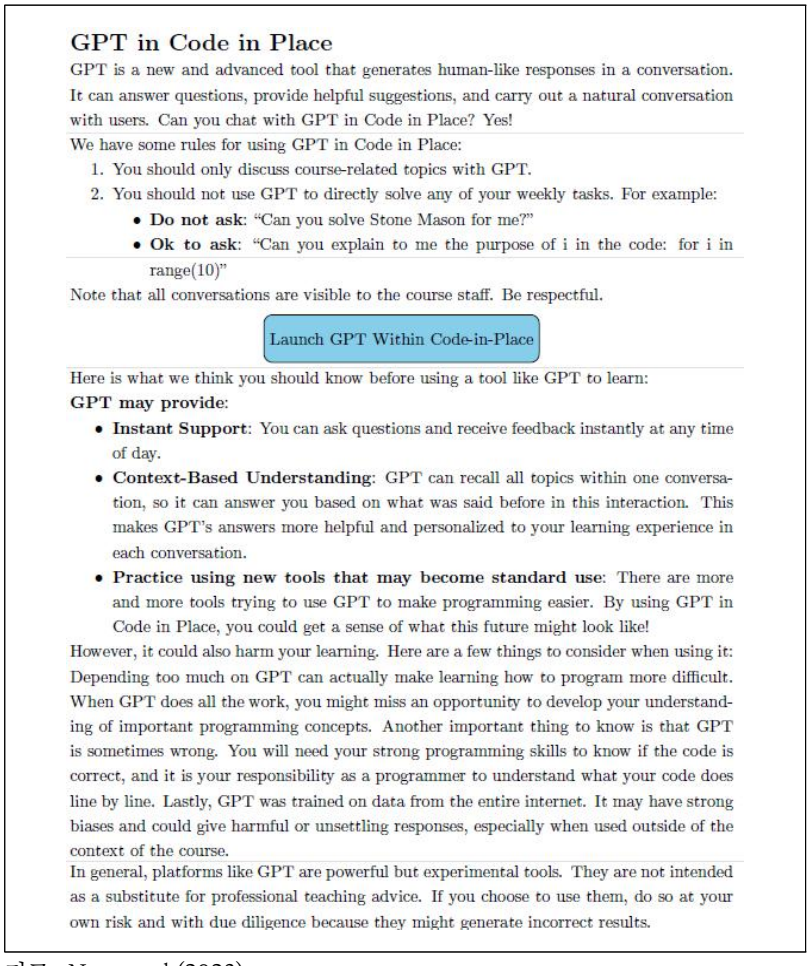
마. Nie et al.(2023)

Nie et al.(2023)은 생성형 AI의 활용이 교육에 미치는 영향을 파악하기 위해 146개국에서 온라인 코딩 교육 강좌에 참여한 5,831명의 학생들을 대상으로 대규모 무작위 대조 실험을 진행하였다. 실험 참가자들 중 생성형 AI를 활용한 학생들은 온라인 코딩 교육 강좌 시험에서 긍정적인 효과를 보는 것으로 나타났으나, 생성형 AI 활용에 대한 광고는 학생들의 시험 참가율을 낮추는 것으로 나타났다. 이는 생성형 AI의 활용이 평균적으로 교육 참여(education engagement)에 부정적인 영향을 줄 수 있음을 시사한다. 하지만 이러한 부정적 효과는 국가별로 상이하였으며, 인적자본지수가 낮은 국가에서는 오히려 교육 참여가 소폭 상승하는 것으로 조사되었다.

Nie et al.(2023)은 코딩 초급자들을 대상으로 한 무료 온라인 강좌를 수강하는 5,831명의 학생들 중 약 60%인 3,581명의 학생들을 무작위로 선택하여 GPT-4의 사용 권한을 부여하였고, 나머지 학생들은 GPT-4를 활용할 권한을 부여하지 않았다. 이러한 실험군과 대조군의 설정은 강좌의 4주 차에 이루어졌고, 5주 차의 마지막 날에는 모든 학생들에게 4시간 정도가 소요되는 중간시험을 공지하였고, 약 45.8%의 학생들이 중간시험에 응시하였다. GPT-4에 대한 활용 권한이 부여된 실험군에서 중간시험 이전에 GPT-4를 활용한 학생들은 약 14.2%인 512명으

로, GPT-4에 대한 활용 권한이 부여된 대부분의 학생들이 GPT-4를 활용하지 않았다.

[그림 2-6] 실험군에게 제시된 GPT-4에 대한 소개



GPT in Code in Place

GPT is a new and advanced tool that generates human-like responses in a conversation. It can answer questions, provide helpful suggestions, and carry out a natural conversation with users. Can you chat with GPT in Code in Place? Yes!

We have some rules for using GPT in Code in Place:

1. You should only discuss course-related topics with GPT.
2. You should not use GPT to directly solve any of your weekly tasks. For example:
   - **Do not ask**: "Can you solve Stone Mason for me?"
   - **Ok to ask**: "Can you explain to me the purpose of i in the code: for i in range(10)"

Note that all conversations are visible to the course staff. Be respectful.

Launch GPT Within Code-in-Place

Here is what we think you should know before using a tool like GPT to learn:

GPT may provide:

- **Instant Support**: You can ask questions and receive feedback instantly at any time of day.
- **Context-Based Understanding**: GPT can recall all topics within one conversation, so it can answer you based on what was said before in this interaction. This makes GPT's answers more helpful and personalized to your learning experience in each conversation.
- **Practice using new tools that may become standard use**: There are more and more tools trying to use GPT to make programming easier. By using GPT in Code in Place, you could get a sense of what this future might look like!

However, it could also harm your learning. Here are a few things to consider when using it: Depending too much on GPT can actually make learning how to program more difficult. When GPT does all the work, you might miss an opportunity to develop your understanding of important programming concepts. Another important thing to know is that GPT is sometimes wrong. You will need your strong programming skills to know if the code is correct, and it is your responsibility as a programmer to understand what your code does line by line. Lastly, GPT was trained on data from the entire internet. It may have strong biases and could give harmful or unsettling responses, especially when used outside of the context of the course.

In general, platforms like GPT are powerful but experimental tools. They are not intended as a substitute for professional teaching advice. If you choose to use them, do so at your own risk and with due diligence because they might generate incorrect results.

자료: Nie et al.(2023)

Nie et al.(2023)은 생성형 AI의 활용이 교육에 미치는 영향을 살펴보기 위하여 두 가지 결과에 집중하였는데, 첫 번째는 학생들의 교육 참여이다. 이는 필수가 아닌 중간시험에 참여하는 학생 비율로 측정하였다. 두 번째는 교육 성과로 이

는 중간시험 성적으로 측정하였다. 실험 결과 GPT-4에 대한 소개 그 자체로 학생들의 교육 참여도는 감소하는 것으로 나타났고, GPT-4를 학습에 활용한 경우, 실험군 중 실제로 활용한 학생들의 성적과 실험군 내부에서 활용하지 않은 학생들의 성적에는 큰 차이가 없으나, 실험군 중 실제로 GPT-4를 활용하여 학습을 진행한 학생들의 성적이 대조군보다는 통계적으로 유의미한 차이를 나타냈다.

GPT-4에 대한 소개로 인한 교육 참여의 부정적인 영향을 분석하기 위해, Nie et al.(2023)은 실험군과 대조군의 중간시험 참여율을 비교하였다. 실험군의 경우 GPT-4에 대한 소개와 함께 해당 생성형 AI를 활용할 수 있는 권한이 부여된 그룹으로, GPT-4에 대한 소개[4]는 GPT-4의 긍정적인 부분(즉각적인 질의응답 등)과 부정적인 측면(Halluciation 가능성, 교육에서의 부정적인 영향 등)을 모두 포함하였다. 실험군의 중간시험 참여율은 44.1%, 대조군의 참여율은 48.5%로 이 둘의 차이는 통계적으로 유의미한 것으로 나타났다. 또한 강좌 6주 차 과제 완료율 역시 대조군이 약 4.6% 더 높은 것으로 나타났는데, GPT-4에 대한 소개가 이루어진 4주 차 이전에는 차이가 없었던 것을 고려할 때, GPT-4에 대한 소개 그 자체가 학생들의 교육 참여에 부정적인 영향을 미칠 가능성을 시사한다. 흥미로운 점은 이러한 부정적 영향이 인적자본지수에 따라 상이하다는 것인데, 인적자본지수가 낮은 국가들에서는 GPT-4에 대한 소개가 교육 참여에 긍정적인 영향을 미치는 것으로 나타났다. 인적자본지수가 낮은 국가들만을 대상으로 실험군과 대조군의 중간시험 참여율을 비교한 결과, 실험군이 42.3%, 대조군이 27.5%로 실험군이 더 높은 참여율을 보였다. 하지만, 해당 표본수가 적어 통계적으로 유의미한 차이를 보이지는 않았다.

두 번째로 GPT-4 활용이 교육 결과에 미치는 영향을 살펴보기 위해 저자들은 중간시험 성적을 비교하였는데, 먼저 실험군 내에서 GPT-4를 활용하지 않은 학

4) GPT-4에 대한 소개 그 자체로 실험 참가자들은 자신들이 향후 실험군이 될 것을 알게 된다는 측면에서 이러한 디자인의 효과를 실험군자각효과(intent-to-treat effect)라고 부르기도 한다.

생들과 대조군의 시험성적을 비교하였다. 이러한 비교는 GPT-4에 대한 소개가 시험성적에 미치는 영향을 측정하는 것으로, GPT-4에 대한 소개 또는 활용에 대한 권한 부여에 따른 교육참여도 하락과 그에 따른 교육결과를 추정해 볼 수 있다. 비교 결과 이 두 집단의 중간시험 성적은 통계적으로 유의미한 차이를 보이지 않았다. 다음으로 실험군 내에서 GPT-4를 학습에 활용한 학생들과 대조군의 시험성적을 비교하였는데, 이러한 비교를 통해 GPT-4 활용이 교육 결과에 미치는 영향을 측정할 수 있다. 비교 결과 GPT-4를 학습에 활용한 학생들의 시험성적이 약 2.89% 더 높으며 이는 통계적으로도 유의미한 차이로 조사되어 GPT-4의 활용이 교육 결과에 긍정적인 영향을 미치는 것으로 나타났다.

마지막으로 인구통계학적 특성 및 국가별 상이성이 학습 시 GPT-4 사용 결정에 미치는 영향을 살펴보았는데, 교육참여도가 높을수록, HDI가 낮을수록, 그리고 나이가 더 많을수록 GPT-4를 학습할 때 더 많이 사용하는 것으로 나타났다.

Nie et al.(2023)의 연구는 큰 표본 수와 교육이라는 흥미로운 주제에 대한 연구로, 다른 연구에서 살펴보지 않았던 생성형 AI의 도입에 따른 '생산성' 측정의 한 측면, 예를 들어 생성형 AI에 대한 접근 권한이 주어졌을 경우, 교육(업무 또는 과업)에 대한 참여도를 살펴보았다는 점에서 의미있는 연구이다. 하지만, 실험참가자들이 온라인 무료 수강 강좌에 참여하고 있기 때문에, 이들이 실험과정에 적극적으로 참여할 유인체계가 제대로 성립되지 않았다. 또한, 실험군과 대조군은 무작위로 나누었으나, 실험군 내에서 GPT-4를 실제로 학습에 활용한 학생들이 해당 강좌에 더 열정적으로 참여하고 있을 가능성이 높기 때문에 이들이 제시하는 GPT-4 활용에 따른 교육 결과에서의 긍정적인 효과가 과대추정되었을 가능성이 있다.

바. Doshi and Hauser(2024)

Doshi and Hauser(2024)는 생성형 AI가 인간의 창의성에 미치는 영향을 측정하기 위하여 온라인 실험을 진행하였다. 실험 결과 생성형 AI는 인간의 창의성

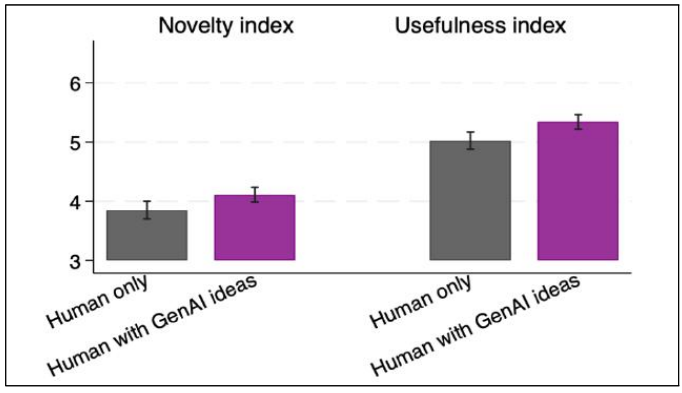
을 높이지만, 생성형 AI의 도움으로 창작된 이야기들은 유사성을 띠는 것으로 나타났다.

Doshi and Hauser(2024)는 2단계로 구성된 온라인 실험을 진행하였다. 총 293명의 참가자들을 모집하였고, 이들에게 8개 문장으로 구성된 짧은 이야기를 작성하는 과업을 부과하였다. 293명의 참가자들은 무작위로 세 그룹으로 나뉘었는데, 세 그룹은 각각 생성형 AI의 도움 없이 과업을 수행하는 그룹(Human Only)과 생성형 AI로부터 이야기의 아이디어를 요청할 수 있는 두 그룹(Human with 1 GenAI idea, Human with 5 GenAI idea)으로 나누었다. 생성형 AI로부터 이야기의 아이디어를 요청하는 방식은 각 참가자들이 이야기를 시작하기 위한 첫 세 문장을 요청하는 것으로, 1 GenAI idea 그룹은 한 번의 요청을, 5 GenAI idea 그룹은 해당 요청을 다섯 번까지 할 수 있다. 이들은 8개 문장으로 구성된 짧은 이야기를 작성한 후 스스로 자신이 작성한 이야기의 참신성(Novelty), 효용성(Usefulness), 그리고 몇몇 감정적 특성에 대하여 평가하는 설문에 응답하였다.

짧은 이야기를 작성하는 293명의 참가자들 이외에 이들의 이야기를 평가하기 위해 총 600여 명의 참가자들을 모집하였으며, 이들은 293명이 작성한 이야기 중 무작위로 선택된 여섯 개의 이야기를 읽고 이들의 참신성(Novelty), 효용성(Usefulness), 그리고 감정적 특성에 대한 평가를 실시하였다. 이야기를 평가할 때는 각 이야기들이 AI의 도움을 받아 작성되었는지에 대한 정보눈 공개되지 않았다.

실험 결과 평가자들은 Human Only 그룹에 비하여 생성형 AI의 도움을 받은 두 그룹의 이야기들의 참신성과 효용성이 더 높은 것으로 평가하였다. 참신성의 경우 생성형 AI의 도움을 받은 이야기들이 평균적으로 약 6.7% 더 높은 점수를 획득하였고, 효용성에서는 6.4% 더 높은 점수를 획득하였으며, 이들은 모두 통계적으로 유의미한 차이를 보였다.

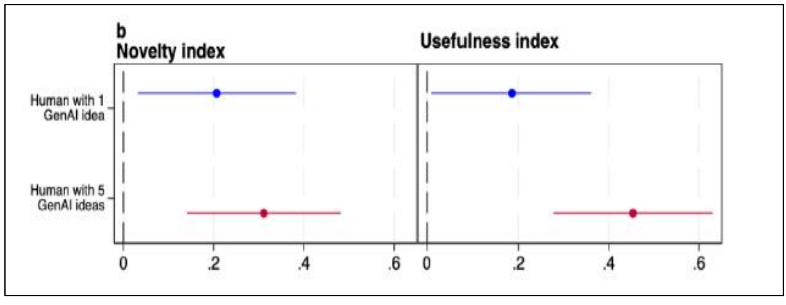
[그림 2-7] 실험 참가 그룹 간 참신성, 효용성 지수 비교



자료: Doshi and Hauser(2024)

다음으로 생성형 AI의 도움을 받은 두 그룹 간의 차이를 비교하였다. 1 GenAI idea 그룹은 총 100명의 참가자 중 약 82명이 생성형 AI의 도움을 받았고, 5 GenAI idea 그룹은 98명 중 93명이 생성형 AI의 도움을 받았다. 참신성과 효용성 두 측면에서 5 GenAI idea 그룹이 1 GenAI idea 그룹보다 높은 점수를 획득하였다. 참신성에서 1 GenAI idea 그룹은 Human Only 그룹 보다 약 5.4% 높은 점수를 받은 반면, 5 GenAI idea 그룹은 약 8.1% 높은 점수를 획득하였다. 효용성에서는 1 GenAI idea 그룹은 Human Only 그룹 보다 약 3.7% 높은 점수를 얻었으나, 5 GenAI idea 그룹은 약 9% 높은 점수를 획득하였다.

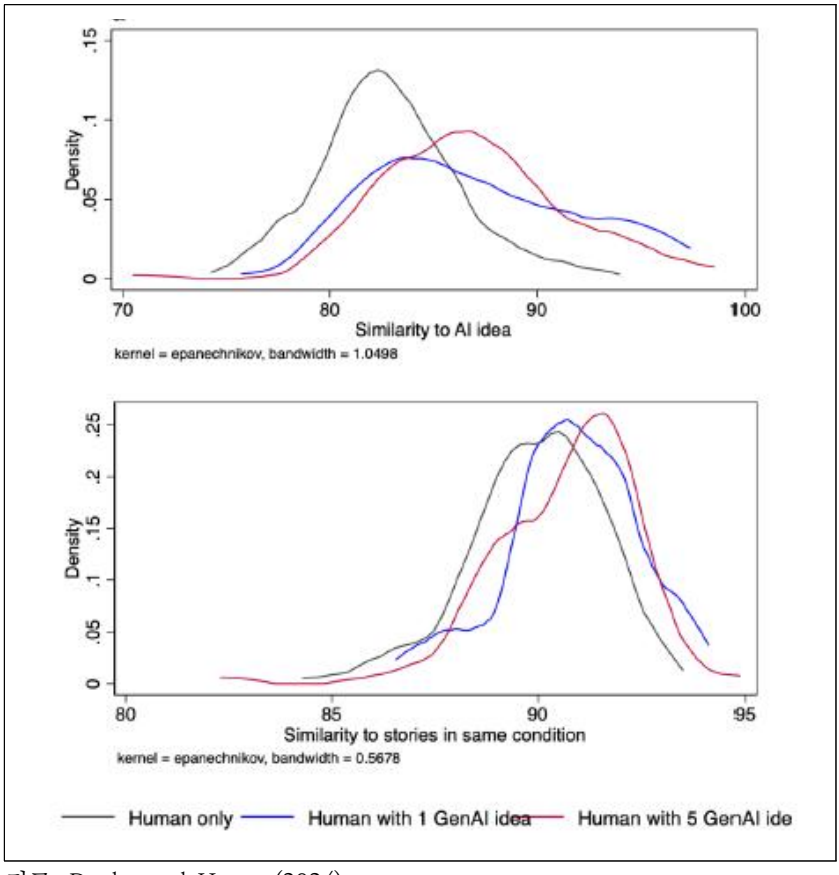
[그림 2-8] 1 GenAI idea 그룹과 5 GenAI idea 그룹 간의 참신성, 효용성 비교



자료: Doshi and Hauser(2024)

마지막으로 이야기의 유사성을 평가하기 위해, 모든 이야기들에 실험 중에 생성형 AI에 의해 제시된 아이디어를 무작위로 지정하고 각 이야기와의 유사성을 평가하였다. 그 결과 Human Only 그룹에 비하여 생성형 AI의 도움을 받은 두 그룹이 작성한 이야기 유사성이 더 높은 것으로 나타났다.

[그림 2-9] 그룹 간 이야기의 유사성 정도 분포



자료: Doshi and Hauser(2024)

사. Haslberger et al.(2024)

Haslberger et al.(2024)는 생성형 AI가 과업과 직무별로 생산성 불평등도를 줄여준다는 연구결과들이 다수를 차지하는 시점에서 이러한 경향을 확인하기 위한 온라인 실험을 수행하였다. 영국 경제활동인구를 대표할 수 있는 표본을 구성하고 이들을 실험군과 대조군으로 무작위 분류하여 실험을 진행하였다. 실험군에게는 ChatGPT의 활용을 장려하고, 대조군에게는 장려하지 않은 상태에서 실제 직무와 비슷한 과업을 수행하도록 하였다. ChatGPT에 노출된 실험군은 모든 과업에서 대조군에 비하여 높은 생산성을 보였고, 특히 복잡도가 높고 덜 모호한 과업에서 생산성 향상이 더 높게 나타났다. 하지만 생산성 불평등도를 낮추는 효과는 제한적으로 관측되었다. ChatGPT는 교육과 직무별로 그룹 간 생산성 차이를 감소시켰으나, 나이대에 따라 나누어진 그룹에서 상대적으로 낮은 연령 그룹과 높은 연령 그룹 간의 생산성 격차는 더 벌어지는 것으로 나타났다.

온라인 실험을 위해 ChatGPT 계정을 소유하고 있는 총 1,041명의 참가자를 모집하였고, 이들 중 무작위로 504명은 ChatGPT 사용을 장려하는 실험군으로, 나머지 참가자들은 ChatGPT 사용을 장려하지 않는 대조군으로 나누었다. 실험군과 대조군 참가자는 총 세 개 과업을 진행하였는데, 이 과업들은 각기 다른 복잡도를 가지고 있다. 실험 참가를 독려하기 위해 상위 10%의 과업 결과물을 제출한 참가자에게는 기본 실험 참여 대가의 50%가 보너스로 지급되었다.

첫 번째 과업은 가상의 직장에서 분쟁을 다루는 이메일이 주어지고 이를 적절하게 수정하는 것으로, 이메일은 의도적으로 문법적 오류와 오탈자를 포함하고 있으며, 비전문적이고 거친 어조로 작성되었다. 첫 번째 과업의 평가는 문법적 오류와 오탈자 수정이 이루어졌는지, 적절치 않은 어조가 조금 더 전문적이고 건설적인 어조로 수정되었는지를 기준으로 진행되었다.

[그림 2-10] 이메일 과업에 주어진 이메일 내용

"Dear Lars,
I am writing to express my disappointment in your recent performance. Your work has not been up to the standard that is being expected from our employers, and it has been impacting the overall productiveness of the team. You lack attention to detail and failure to meet deadlines has resulted in missed opportunities and last revenue for the company. This is unacceptable and cannot continued.
I want to make it clear that if your performance does nor improve significantly in the coming weeks, we will have no choice but to conclude your employment with us. We have highest standards here, and we expect all employees to meet them.
I strongly encourage you to take a clear look at your life and make the necessary changes to improve. This includes better time management, increased attention to detail, and increased communications with your colleagues. We are here to support you, but ultimately it is up to you to make the initiative to improve your work.
I hope to see a significant improvement of your performance in the future weeks. If you have any questions or concerns, please do not hesitate to reach us.
Sincerely,
Restaurant Owner"

자료: Haslberger et al.(2024)

두 번째 과업은 500단어로 구성된 두 문서의 설득력을 평가하는 것인데, 각 문서는 보편소득에 대해 찬성하고 반대하는 글로 구성되었다. 참가자들은 이 두 문서 중 하나를 선택하고 그 문서의 주장이 어떠한 이유에서 더 설득력이 있는지를 평가하여 답하여야 한다. 해당 과업의 평가는 둘 중 하나의 문서를 선택한 참가자가 그 이유를 얼마나 논리적으로 서술하는지에 따라 이루어졌다.

마지막 과업은 아일랜드의 경제사에 관한 복잡한 문서를 바탕으로 세 가지 짧은 질문에 답하는 것이다. 평가 방법은 참가자들이 복잡한 문서에서 얼마나 잘 정보를 정제해내는지로 진행되었다.

이들 세 과업은 생성형 AI의 도움 없이는 점차적으로 그 복잡도가 높아지는 특성을 가지도록 고안되었고, 대다수의 사람들이 이미 가지고 있는 사전지식이나 경험으로는 쉽게 해결할 수 없는 과업들로 구성되었다.

세 과업은 GPT-4를 활용하여 평가하였는데, GPT-4의 평가가 내생적으로 편향되었을 가능성을 검사하기 위해, 100개의 표본을 구성하여 동일한 기준으로 다른 생성형 AI를 활용하여 평가한 결과, 결과 값에 큰 차이가 나타나지 않았다. 두

번째 평가척도는 각 과업을 완료한 시간으로 과업당 참가자들이 최대로 사용할 수 있는 시간은 900초로 제한되었다. 평가점수와 시간을 활용하여 실험군과 대조군의 생산성 차이를 식별하기 위해, 실험군과 대조군을 표시하기 위한 지시함수(Indicator function)와 인구통계학적 변수들을 포함한 간단한 회귀분석을 실시하였다.

분석 결과 모든 과업에서 ChatGPT 활용을 장려받은 그룹이 대조군보다 더 빨리 과업을 완료하였을 뿐만 아니라 더 높은 평가점수를 기록하였다. 이메일 과업의 경우 실험군이 약 90초 더 빠르게 완료하였고, 5점 척도 기준으로 약 0.8점 더 높은 평가를 받았다. 두 번째 과업에서는 실험군이 약 170초 빠르게 과업을 완료하였고, 약 0.4점 높은 점수를, 세 번째 과업에서 역시 실험군이 약 60초 더 빨리 과업을 완료하였으며 평가점수는 약 1.3점 높았다.

마지막으로 생성형 AI 활용이 집단 또는 직무별 생산성 격차를 줄이는지 알아보기 위하여, 각 과업별로 생산성을 나타내는 계수의 편차를 활용하였는데, 만약 생성형 AI 활용이 생산성 격차를 줄인다면, 실험군에서 생산성 관련 계수의 편차가 더 낮게 나타나야 한다. 하지만 분석 결과, 생산성 계수의 편차를 실험군과 대조군을 각각 비교한 결과, 생산성 격차를 줄인다는 증거는 발견할 수 없었다. 즉, 이메일 과업의 경우 실험군과 대조군의 차이가 미미하였으며, 두 번째 과업에서는 실험군 내에서의 생산성 계수의 편차가 대조군보다 더 높게 나타났으며, 세 번째 과업에서만 생산성 격차를 줄이는 형태로 실험군에서의 생산성 계수의 편차가 대조군보다 작게 나타났다.

아. Jia et al.(2024)

Jia et al.(2024)은 텔레마케팅 회사에 근무하는 직원을 대상으로 현장실험(field experiment)을 진행하였다. 실험 결과 잠재적 고객을 확보하는 과정에서 생성형 AI를 활용하는 것이 직원들이 추후에 구매를 설득할 때 고객들과 질의응답에서 창의성을 높이는 것으로 나타났다.

현장실험을 위해서 Jia et al.(2024)은 아시아에 위치한 텔레마케팅 회사와 함께 연구를 진행했다. 해당 회사는 3,000만 명이 넘는 고객을 대상으로 다양한 종류의 상품을 판매하는 것에 특화되어 있다. 현장실험이 진행될 당시 해당 회사는 새로운 사업으로 신용카드 판매를 주요 은행과 함께 진행할 예정이었고, 따라서 해당 회사의 직원들은 그 누구도 신용카드를 판매한 경험이 없는 상태였다.

해당 회사는 판매 업무를 두 개의 과업으로 나누어 진행했는데, 첫 번째 단계에서 직원들은 고객에게 전화를 걸어 판매 상품에 대한 정보를 제공하고 해당 고객이 상품에 관심이 있는지를 알아보았다. 이 단계에서 관심을 표하지 않은 고객은 걸러지고, 관심을 표한 고객을 대상으로 두 번째 과업이 진행되었다. 이러한 과업은 해당 회사에서 상당히 잘 구조화된 과업으로 많은 대응 방식과 대본이 이미 작성되어 있었다. 두 번째 과업은 구매설득으로 관심을 표한 고객의 필요를 알아내고 해당 상품이 고객의 필요를 충족시켜 줄 수 있음을 설득함으로써 구매결정을 이끌어내는 설득의 과정이었다. 두 번째 과업은 첫 번째 과업에 비해 덜 구조화되어 있다.

현장실험에서는 업무성과를 기준으로 상위 33%의 직원 20명과 하위 33% 직원 20명을 모집한 후 이들을 생성형 AI의 지원을 받는 실험군과 그렇지 않은 대조군으로 무작위로 나누어 실제 업무를 진행하게 하였다. 생성형 AI는 첫 번째 과업에서 활용되어, 실험군의 경우 첫 번째 과업은 생성형 AI가, 두 번째 과업은 직원이 수행하는 반면, 대조군은 두 과업을 모두 직원이 수행하였다.

실험참가자들의 과업 평가 기준은 두 과업에서의 창의성과 성과인데, 창의성은 구매설득 과정에서 사전에 교육받지 못한 판매상품(신용카드)에 대한 질문을 해결한 횟수로 측정하였고, 성과는 구매결정을 내린 고객의 수로 측정하였다.

실험 분석 결과, 먼저 생성형 AI의 지원을 받은 직원의 창의성이 대조군에 비하여 통계적으로 유의미하게 더 높았다. 즉, 생성형 AI의 지원을 받은 실험군의 직원들이 사전에 교육받지 않은 제품에 대한 질문을 약 2.33배 더 많이 성공적으로 해결하였다. 상위 33%의 직원들과 하위 33% 직원들로 나누어 생성형 AI의 생산성 향상을 비교한 결과, 상위 33%에서는 실험군과 대조군의 차이가 하위 33%에

서의 실험군과 대조군 차이의 약 2.8배로 나타나 숙련된 직원들이 생성형 AI의 지원을 받을 경우 비숙련 직원들보다 더 큰 생산성 향상을 보였다. 또한 성과 측면에서도 생성형 AI의 지원을 받은 실험군이 대조군보다 약 60% 더 높은 평가를 받았다.

텔레마케터라는 직군에 집중하여 실제 업무를 통해 현장실험을 진행하였다는 점에서 매우 중요한 연구이지만, 표본수가 40으로 매우 적고 연구에서 활용하는 창의성과 성과라는 척도에 따른 평가 결과가 매우 작은 숫자로 나타나 이들의 통계적 유의성에 대한 논의가 필요하다.

자. Roldan-Mones(2024)

Roldan-Mones(2024)는 대학교의 토론대회 참가자들을 대상으로 한 무작위 대조 실험(Randomized Controlled Trial)을 진행하여 생성형 AI의 생산성을 측정하였다. 생성형 AI를 활용하여 진행되는 과업이 기존 문헌에서 활용되었던 과업에 비하여 높은 수준의 인지 및 사회적 상호작용 기술, 그리고 예측할 수 없는 언어적 소통을 포함한다는 점에서 이전의 연구와의 차별성을 가진다. 실험 결과 상위권 학생들이 그렇지 않은 학생들보다 생성형 AI의 활용으로부터 더 큰 효용을 얻는 것으로 나타났다. 이는 생성형 AI의 활용이 숙련/비숙련 노동자 또는 고급/일반 인력의 생산성 차이를 줄여준다는 기존의 연구 결과와 상반되는 것으로, 상위권 학생들이 생성형 AI로부터 제공받은 정보들을 더 효과적으로 사용하기 때문으로 추측된다. 기존 연구 결과와 해당 연구 결과의 차이점은 과업의 차이에서 기인하는 것으로, 과업이 단순 반복 또는 예상 가능한 범주에서의 변화만을 포함하지 않을 경우, 숙련 또는 고급 인력에서 생성형 AI의 생산성 향상이 더 크게 나타날 수 있음을 시사한다.

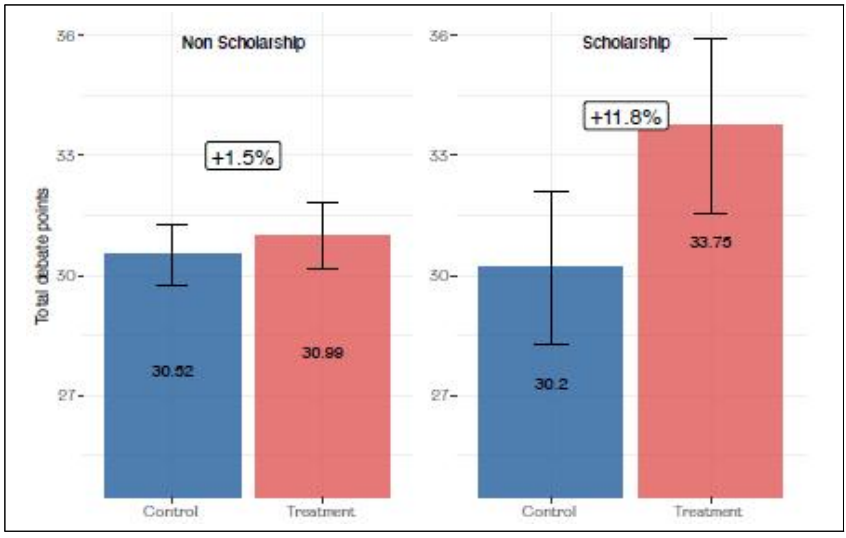
실험은 ESADE 경영학부 학생들과 CUNEF 대학교의 경제정책 과목을 수강하는 학생들을 대상으로 이루어졌다. 총 세 개의 토론 세션에서 실험이 진행되었는데, 첫 두 개의 세션은 ESADE 경영학부 학생들을 대상으로 각기 다른 날에 진행

되었고, 첫 세션에서는 38명, 두 번째 세션에서는 50명의 학생이 참가하였다. 세 번째 세션은 CUNEF 대학교의 경제정책 과목 수강생을 대상으로 진행되어 총 58명의 학생이 참가하였다. 각 세션은 3시간 동안 진행되었으며, 각 세션에서 3~4번의 일대일 토론이 공공정책과 관련된 주제로 진행되었다. 각 토론은 토론자 1인당 약 5분의 시간이 주어졌는데, 5분 동안 각 토론자에게는 3분의 모두 발언과 2분의 반박 발언 시간이 주어졌다. 토론 내용은 모두 녹음되어 토론 평가자들에게 제공된 후 평가되었다. 평가자들은 토론대회 우승자이거나 대학교에서 토론 과목을 강의하는 강사들로 구성되었다. 총 9명의 평가자를 섭외하였고, 각 평가자에게 무작위로 토론이 배당되어, 5가지 평가기준별로 각각 10점 만점 척도로 평가가 진행되었다. 5가지 평가기준은 ① 방어논리의 명료성과 유효성, ② 사용된 증거의 신뢰성, ③ 화법의 적절성, ④ 반대 진영의 논리를 반박할 수 있는 능력, ⑤ 사용한 주장의 우월성이다. 토론자들의 적극적인 참여를 유도하기 위하여, 총 10개의 100유로 상당의 아마존 상품권이 상품으로 주어졌다. 이 중 5개 상품은 대조군(생성형 AI를 활용하지 않은 그룹)에게, 5개 상품은 실험군에게 주어졌다.

먼저 참가자들은 무작위로 실험군과 대조군으로 분류되었다. 그 후 각 토론별로 만나게 되는 상대방 역시 무작위로 배정되었고, 마지막으로 토론에서의 찬반 입장 역시 무작위로 배정되었다.

모든 참가자들은 노트북(Laptop)과 핸드폰을 소지한 채로 토론에 임하였는데, 첫 번째 토론은 모든 참가자들에게 ChatGPT의 사용을 금지한 상태에서 진행되었다. 첫 번째 토론이 끝난 후 실험군은 약 20분간의 ChatGPT 활용 교육을 받았고, 대조군은 20분간 휴식을 취하였다. 실험군과 대조군은 각기 다른 방에서 이후의 토론 주제를 공지받은 후 약 45분간 토론을 준비할 시간을 가진 후 두세 번의 토론을 진행하였다.

[그림 2-11] 장학생 그룹과 비장학생 그룹의 생산성 향상 효과 비교



자료: Roldan-Mones(2024)

분석을 위하여 간단한 OLS 회귀분석이 사용되었다. 먼저 실험군과 대조군의 토론 성과를 비교한 결과, 실험군이 토론 점수와 토론 승률이 대조군보다 높게 나타났으나 통계적으로 유의미한 차이를 보이진 않았다. ChatGPT의 생산성 향상 효과를 장학생과 그렇지 않은 학생으로 나누어 살펴본 결과, 장학생의 경우 ChatGPT를 활용한 실험군의 토론 점수가 대조군보다 약 11.8% 높게 나타났고 통계적으로 유의미한 차이를 보인 반면, 비장학생의 경우 실험군이 약 1.5% 높은 점수를 기록하였으나 통계적으로 유의미한 차이는 아닌 것으로 나타났다.

차. Kim and Moon(2025)

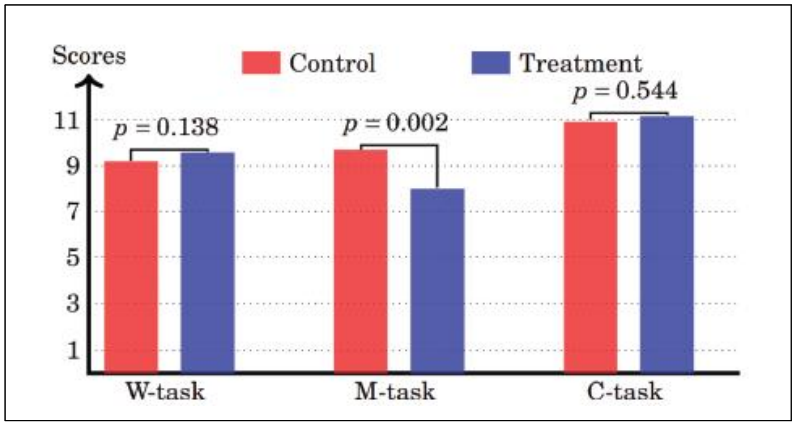
Kim and Moon(2025)은 대학생들을 대상으로 ChatGPT가 읽기와 쓰기, 수학 문제 풀이에 도움이 되고 코딩 관련 사고를 증진시키는지를 실험을 통하여 분석하였다. 분석 결과, 읽기와 쓰기 그리고 코딩 관련 사고 과업에서는 ChatGPT가 과업성과를 유의미하게 변화시키지 못한 반면, 수학문제 풀이에서는 ChatGPT를

활용한 실험군의 성과가 대조군에 비하여 유의미하게 감소하는 것으로 나타났다. 이러한 결과는 기존 문헌에서 보고된 추세와 반대되는 것으로, 수학문제 풀이의 경우 ChatGPT를 활용한 참가자들이 Hallucination으로 인해 발생한 오답을 거르지 못함으로써 오답률이 증가한 것으로 추정하였다.

실험은 한국의 한 대학교에서 만 19세 이상 349명의 학생들을 대상으로 진행되었다. 349명의 학생들은 무작위로 실험군(과업 수행에서 ChatGPT를 활용하는 그룹)과 대조군으로 나누었다. 각 그룹에 속한 참가자들은 무작위로 읽기와 쓰기, 수학문제 풀이, 코딩 관련 사고능력으로 구성된 총 세 개의 과업 중 두 개를 수행하였다. 실험군은 ChatGPT를 사용하도록 장려하였으며, 이들은 과업 실시 이전에 ChatGPT 사용에 익숙해지는 시간을 가졌으며, 대조군은 ChatGPT를 제외한 모든 기능을 활용할 수 있는 상황에서 과업을 수행하였다. 참가자들은 과업의 성과에 따라 실험 참가의 대가를 받았는데, 최소는 기본참여 대가인 13,500원, 가장 높은 대가는 27,000원이었다.

실험 결과 읽기와 쓰기 그리고 코딩 관련 사고능력 과업에서는 실험군과 대조군의 성과지표에서 통계적으로 유의미한 차이가 존재하지 않았다. 반면 수학문제 풀이에서는 실험군의 점수가 대조군에 비하여 통계적으로 유의미한 수준에서 더 낮게 나타났다. 이는 ChatGPT가 Hallucination에 의해 제시한 오답을 참가자들이 걸러내지 못한 것으로 추정된다.

[그림 2-12] 실험 결과



자료: Kim and Moon(2025)

카. Dell'Aqua et al.(2024)

Dell'Aqua et al.(2024)은 보스턴컨설팅 그룹의 컨설턴트 758명을 대상으로 생성형 AI의 활용이 복잡하고 지식집약적인 과업에 미치는 영향을 분석하기 위하여 실험을 진행하였다. 실험참가자들은 무작위로 생성형 AI를 활용하지 않는 그룹, GPT-4에 대한 접근이 가능한 그룹, 그리고 GPT-4에 대한 접근과 프롬프트 엔지니어링에 대한 개략적인 교육을 받은 그룹으로 나누었다. 분석 결과 생성형 AI가 쉽게 처리할 수 있는 과업과 난이도는 비슷하지만 현재 생성형 AI로는 큰 도움을 받지 못하는 과업이 있는 것을 확인하였다. 총 18가지 컨설팅 과업이 활용되었고, 생성형 AI를 활용하는 그룹의 생산성 향상을 확인하였는데, 생성형 AI를 활용한 그룹이 그렇지 않은 그룹에 비하여 12.2% 더 많은 과업을 완료하였으며, 약 25.1% 더 빨리 과업을 완료하였다. 업무효율성이 높은 참가자들의 경우 17%의 생산성 향상이 나타난 반면 그렇지 않은 참가자들의 경우 43%의 생산성 향상이 나타나, 기존 문헌에서 보이는 생성형 AI의 활용이 업무효율성에 따른 상이한 생산성 격차를 줄이는 결과를 다시 한번 확인하였다.

실험은 보스턴컨설팅 그룹의 컨설턴트들을 대상으로 진행되었다. 실험은 약 5

시간 동안 진행되었으며 약 7%의 보스턴컨설팅 그룹의 컨트리뷰터 컨설턴트 집단이 실험에 참가하였다. 실험은 총 3단계로 진행되었는데, 첫 번째 단계에서 참가자들은 인구통계학적 정보와 심리학적 프로파일에 대한 설문에 응답하였다. 몇 주 후에 참가자들은 실제 실험에 참가하기 위한 온라인 링크를 수령받고 과업 이전 설문, 실제 과업 수행, 그리고 과업 완료 후 설문이 진행되었다. 마지막 단계에서는 참가자들을 대상으로 그들이 가진 생성형 AI에 대한 의견과 실제 경험에 대한 인터뷰가 진행되었다,

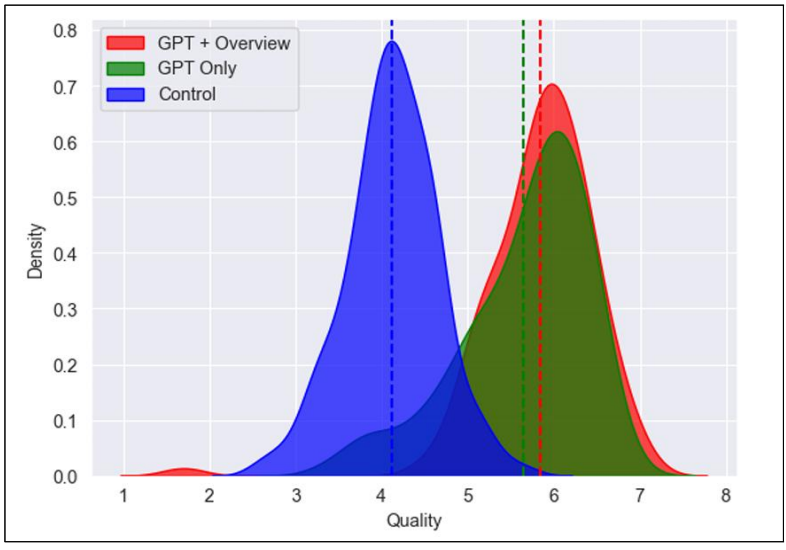
실제 실험에서는 두 종류의 과업에 따라 총 738명의 참가자들을 다시 두 그룹으로 나누었는데, 첫 번째 종류의 과업은 새로운 제품에 대한 아이디어를 개발하는 것으로 창의성, 분석능력, 설득력, 글쓰기와 같은 능력에 초점을 맞춘 과업으로, 약 385명의 참가자들이 해당 과업을 수행하였다. 두 번째 종류는 기업경영에서 발생할 수 있는 문제를 해결하는 과업으로 데이터 분석, 소비자 및 기업 인터뷰, 설득력, 글쓰기와 같은 능력이 유용하게 사용되는 과업으로 약 373명의 참가자들이 해당 과업을 수행하였다.

각 과업별로 참가자들은 먼저 생성형 AI를 활용하지 않은 상태에서 과업을 수행하고 이러한 과업 결과는 생성형 AI를 활용하여 수행한 과업의 결과물과 비교하기 위한 기준점(Baseline)으로 활용하였다. 첫 번째 과업을 마친 참가자들은 다시 무작위로 AI를 활용하지 않는 그룹, AI를 활용하지만 프롬프트 엔지니어링에 대한 개략적인 내용은 활용할 수 없는 그룹, AI와 프롬프트 엔지니어링에 대한 개략적인 내용을 모두 활용할 수 있는 그룹으로 나누어 다시 과업을 수행하는 형태로 실험이 진행되었다. 과업 결과는 두 명의 전문가가 평가하였으며, GPT-4를 활용하여 각 과업 결과물을 다시 평가하였다.

먼저 생성형 AI가 쉽게 처리할 수 있는 과업에 관한 실험은 틈새시장 공략을 위한 신발 콘셉트를 구상하는 것으로 프로토타입 묘사부터 시장 층위 분석과 시장 진입까지의 모든 단계에 대한 자세한 서술을 요구한다. 실험 결과, 생성형 AI를 활용한 그룹이 그렇지 않은 그룹에 비하여 과업 결과 평가점수가 평균적으로 1.75점 더 높게 나타나 약 42.5%의 생산성 증가를 나타냈다. 또한 과업완료 비율

역시 생성형 AI를 활용한 그룹이 그렇지 않은 그룹보다 12.2% 더 높은 것으로 나타났다.

[그림 2-13] 평가점수 분포 비교

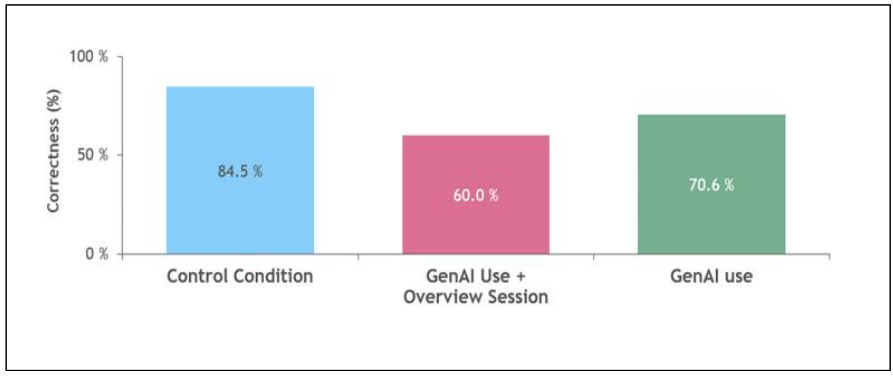


자료: Dell'Aqua et al.(2024)

다음으로 첫 번째 실험의 과업과 비슷한 난이도를 가지고 있지만, 생성형 AI가 쉽게 해결할 수 없는 과업을 대상으로 실험이 진행되었다. 해당 실험의 과업은 가상 회사의 유통 채널 분석부터 브랜드 분석을 바탕으로 가상 회사의 CEO에게 어떤 브랜드가 가장 높은 잠재력이 있는지를 추천하는 과업으로 참가자들은 인터뷰 정보, 재무정보 등을 활용하여 해당 과업을 수행하였다. 이러한 종류의 과업은 생성형 AI가 단순히 정보를 찾고 분석하는 것으로는 완료할 수 없는 과업이다. 해당 과업의 평가가 합리적인 판단인가 그렇지 않은 판단인가를 바탕으로 이루어졌다.

실험 결과 생성형 AI를 활용한 그룹이 그렇지 않은 그룹에 비하여 올바른 판단을 내린 비율이 더 낮은 것으로 나타나, 생성형 AI 활용에 따른 부정적인 영향이 확인되었다.

[그림 2-14] 과업 평가결과 비교



자료: Dell'Aqua et al.(2024)

〈표 2-1〉 선행 연구 요약

논문(저자)	실험분석방법	실험대상자	과업	생산성 변화			과업평가방법	참가자 유인 제공 여부
Noy and Zhang(2023)	그룹 간 비교(Between)/ 그룹 내 비교(Within)	대졸 전문직 (n=453)	실제 업무 (상품 보도자료, 데이터 분석, 제안서 작성, 이메일 작성, 보고서 작성)	구분	그룹 간 비교	그룹 내 비교	동종업계 전문가 평가	높은 점수일수록 높은 대가 지급
				속도	-42% (-16분)	-30% (-10분)		
				평가 점수	+15% (+0.45점)	+26~56% (+1-2점)		
Choi and Schwarcs (2023)	그룹 간 비교(Between)/ 그룹 내 비교(Within)	학생들 (n=48)	실제 시험	구분	그룹 간 비교	그룹 내 비교	시험성적	없음
				시험 성적	+29%	-		
				높은 생산성 집단	-	-20%		
				낮은 생산성 집단	-	+40%		
Brynjolfsson et al.(2023)	그룹 간 비교(Between)	고객상담센터 근무자 (n=5,172)	고객상담업무	구분	그룹 간 비교	그룹 내 비교	상담건수 및 상담완료소요시간	없음 (실제 업무에서 활용)
				상담완료 건수/1시간	+15% (+0.3건)	-		
				소요시간/상담1건	-9% (-3.7분)	-		
Peng et al. (2023)	그룹 간 비교(Between)	전문 개발자 (n=95)	코딩	구분	그룹 간 비교	그룹 내 비교	과업완료시간	과업완료시간이 빠를수록 높은 대가 지급
				완료시간	-55.8%	-		

논문(저자)	실험분석방법	실험대상자	과업	생산성 변화			과업평가방법	참가자 유인 제공 여부
Nie et al. (2023)	그룹 간 비교(Between)	온라인 코딩 무료강좌 수강 학생(n=5,831)	중간시험	구분	그룹 간 비교	그룹 내 비교	시험성적	없음
				시험성적	+2.89%	-		
Doshi and Hauser(2024)	그룹 간 비교(Between)	참가자(n=293) 평가자(n=600)	이야기 작성	구분	그룹 간 비교	그룹 내 비교	평가자들의 참신성 및 효용성 평가	없음[5]
				참신성	+6.7%	-		
				효용성	+6.4%	-		
Haslberger et al.(2024)	그룹 간 비교(Between)	ChatGPT 계정보유자 (n=1,041)	이메일 수정, 설득력 평가, 문서 이해	구분	그룹 간 비교		과업완료 속도 및 점수 (GTP가 평가)	상위 10%에게 보너스 지급[6]
					과업완료 속도	평가점수		
				이메일 수정	-30%	+26%		
				설득력	-54%	+18%		
				문서 이해	-20%	+67%		
Jia et al. (2024)	그룹 간 비교(Between)	텔레마케팅 회사 직원 (n=40)	제품 판매	구분	그룹 간 비교		창의성 (질문해결 건수) 성과 (제품판매 건수)	실제 업무에 필요하지 않음
				창의성	+130%			
				성과	+110%			
Roldan-Mones (2024)	그룹 간 비교(Between)	토론대회 참가 학생 (n=146)	정책 이슈에 관한 토론	구분	그룹 간 비교		토론교육 전문가의 평가점수	상품(100유로 상당의 아마존 상품권)
				전체	유의한 차이 없음			
				장학생	+11.8%			
				비장학생	유의한 차이 없음			

5) 기본대가는 지급되었으나, 과업에 최선을 다할 유인은 제공되지 않음

6) 다만 이러한 유인체계가 대조군이 ChatGPT를 활용할 여지를 준다는 점에서 한계가 존재

논문(저자)	실험분석방법	실험대상자	과업	생산성 변화			과업평가방법	참가자 유인 제공 여부
Kim and Moon (2025)	그룹 간 비교(Between)	대학생(n=349)	읽기 쓰기, 수학문제 해결, 코딩 관련 사고	구분	그룹 간 비교		과업별 점수	높은 점수일수록 보너스 대가 지급
				읽기/쓰기	유의한 차이 없음			
				수학문제	-12.5%			
				코딩사고	유의한 차이 없음			
Dell'Aqua et al.(2024)	그룹 간 비교(Between)	BCG 컨설턴드 (n=758)	실제 과업과 유사한 과업 (상품 프로모션, 브랜드 평가)	구분	그룹 간 비교		전문가 평가 및 GPT 활용 평가	높은 점수일수록 인사고과 점수에 보너스 점수 부여
					과업완료율	평가점수		
				상품 프로모션	+12.2%	+42.5%		
				브랜드 평가	-23%			

자료: 저자 작성

제 3 장 생성형 AI의 생산성 분석

제 1 절 실험설계

1. 실험 설계 전 고려 사항

앞의 문헌에서 살펴본 것처럼, 2023년에서 2024년까지 약 2년 동안 생성형 AI의 생산성 분석을 위한 실험연구들이 많이 진행되었다. 이들 연구들을 살펴보면 실제 업무와 비슷한 과업이거나 대부분 문서 작성 또는 문제 풀이와 관련된 과업을 실험에 활용하였다. 대표적인 생성형 AI인 ChatGPT가 문자로 이루어진 결과물을 제공한다는 점에서 문자로 작성된이루어진 과업이 실험에 사용되는 것은 당연하다. 하지만 최근에는 미드저니 등 문자로 이루어진 과업뿐만 아니라 예술 직군에서도 생성형 AI가 사용되고 있기 때문에 이와 관련된 실험 역시 고려해 볼 만하다.

기존 문헌들에서는 대조군과 실험군을 나누고 이들의 최종 결과물의 점수, 완료 시간을 비교함으로써 생산성 향상을 측정하고 있다. 하지만 대부분의 연구가 왜 이러한 생산성 향상이 일어나는지에 대해서는 구체적인 설명이 없거나, 설명이 존재하더라도 데이터 등 객관적인 증거가 뒷받침되고 있지 않다. 따라서 실험이라는 방법론을 활용하여 생산성 증가를 일으키는 경로를 파악하는 것도 기존 연구에 더하여 의미있는 연구가 될 것이다.

마지막으로 기존 문헌들의 실험이 실제로 진행된 시기에는 생성형 AI의 사용이 대중화되기 이전으로, 실험군과 대조군을 나누고 실험군에게만 생성형 AI에 대한 접근을 허용함으로써 대조군이 생성형 AI의 유료서비스에 접근하지 못하도록 하는 형태의 제약을 실제로 실행하는 데에 큰 문제가 없었다. 하지만 최근에는 생성형 AI의 사용이 많이 대중화되어, 대조군에게 생성형 AI활용을 금지하도록 하는

것이 더욱 까다로워졌다. 이상적인 실험을 위해서는 대조군에 속하는 참가자들에게 생성형 AI를 활용할 수 없는 노트북이나 컴퓨터를 지급하고 실험을 진행해야 하지만, 최근 단순한 구글 검색 역시 생성형 AI인 Gemini와 연동되어 생성형 AI의 사용을 강제하기가 힘들다. 따라서 이러한 환경을 고려한 실험을 고안해야 할 것이다.

2. 실험 설계

본 연구에서는 생성형 AI의 생산성 향상 경로를 파악하기 위한 실험을 설계하고 진행하고자 한다.[7] 본 연구에서 초점을 맞춘 생산성 향상 경로는 AI의 활용과 분업 간의 상호작용에 기반한다. 실험에서 수행될 과업은 과업1과 과업2로 이루어져 있으며, 과업2를 수행하기 위해서는 과업1의 결과물이 필요하다. 따라서 상황에 따라 과업 수행에서 한 사람이 과업1과 과업2를 모두 순차적으로 수행하거나, 과업1을 한 사람이 먼저 수행한 뒤, 또 다른 한 사람이 과업1의 결과물을 전달받고 과업2를 수행하는 형태로 분업이 이루어질 수 있다. 이러한 과업에서 만약 생성형 AI가 활용된다면, 먼저 분업 자체가 소멸될 수 있다. 즉, 이전에는 과업1과 과업2를 동시에 수행하는 데에 많은 시간이 필요하였지만, 생성형 AI를 활용하면 두 과업의 난이도가 낮아지고, 결국 분업의 필요성이 낮아져 분업이 사라질 수 있다. Smith(1776)에 따르면, 분업은 전통적으로 생산성 향상의 원천이다. Smith(1776)의 경우, 반복된 작업에 의한 숙련도 향상, 작업 전환 비용의 감소 등으로 분업이 이루어지면 생산성이 향상될 수 있다고 보았다. 하지만, 분업이 반드시 생산성 향상에 도움이 되는 것은 아니다. 예를 들어, 분업에서 과업1을 수행하는 사람이 과업2를 수행하는 사람의 상황을 완전히 내재화할 수 없다. 즉, 과업

7) 예술 직군의 생산성 향상 여부의 측정 역시 의미있지만, 실제 실험 진행에서 예술 직군 실무자들의 과업 결과물(그림 또는 일러스트)에 대한 소유권 문제, 예산 제약, 샘플 수 확보 등을 고려할 때 실험 진행이 어려웠다. 특히 예술 직군 중 실제 실무자의 경우, 3시간 정도의 실험에 이들의 일일 급여의 3시간에 해당하는 기본답례만으로도 상당한 예산이 소요되는 것으로 예상되었고, 확보 가능한 샘플 수에도 명백한 한계가 존재하였다.

1을 수행하는 사람은 뒤이어 과업2를 수행하는 사람에게 필요한 정보를 완전히 예상할 수 없으며, 과업2 수행을 위해 가장 효율적인 형태의 과업1 결과물 역시 개인 차가 존재하므로 예상할 수 없고, 과업1의 결과물을 이해하는 데 시간이 소요되거나 잘못 이해할 수 있다. 이들 모두 정보손실비용 또는 해석비용으로 해석될 수 있는데, 이러한 비용이 분업의 생산성 향상 효과보다 더 크다면 분업이 곧 생산성 하락으로 귀결될 수도 있다(Radner, 1993; Garicano, 2000; Dessein and Santos, 2006; Dessein, Galeotti and Santos, 2016) 이러한 경우 생성형 AI 활용으로 인해 분업이 소멸되고, 생산성이 향상될 수 있다.

본 연구에서는 생성형 AI의 활용이 분업과 어떠한 형태로 상호작용하며 생산성을 향상시키는지 살펴보기 위하여 실험을 설계하고 진행하였다. 먼저 실험에서는 마케팅 관련 과업을 선택하였다.[8] 실험에 활용된 과업은 새로운 제품에 대한 1,200자 분량의 설명문에 포함된 정보를 바탕으로 구매설득 이메일을 발송할 마케팅 층(Marketing Segment)을 설정하고, 구매설득 이메일을 발송하는 것이다. 이 과업은 총 두 개의 과업으로 구성되는데, 먼저 과업1은 1,200자 분량의 설명문을 읽고, 해당 설명문을 새로운 제품의 기능 세 가지, 구매유도 요인(잠재적 구매자에게 구매를 설득할 수 있는 장점) 세 가지, 그리고 구매저해 요인(잠재적 구매자가 구매를 꺼릴 수 있는 단점) 세 가지를 요약하는 것이다. 다음으로 과업2는 과업1의 결과물을 바탕으로, 구매설득 이메일을 발송할 잠재적 구매층을 연령, 성별, 직업별로 특정하고 설정된 잠재적 구매층을 대상으로 구매설득 이메일을 작성하는 것이다. 과업에서 사용된 새로운 제품은 현실에는 존재하지 않는 제품으로 ChatGPT를 활용하여 위에 언급한 정보들이 포함되도록 1,200자 분량으로 작성하였다.[9] 또한, 과업1과 과업2의 분리는 Dell'Aqua et al.(2024)의 연구를 참조

8) McKinsey(2023)에서 살펴본 바와 같이, 생성형 AI의 경제적 효과가 가장 큰 직무는 마케팅 및 제품 판매이므로, 마케팅 과업의 경우 생성형 AI의 생산성 향상 효과가 어느 정도 확인되었기에, AI 활용군의 처치효과가 명확해질 것이다.

9) 실제 존재하는 제품의 경우, 실험참가자가 이미 정보를 가지고 있을 가능성이 있어 실험의 객관성을 높이기 위해 가상의 제품을 상정하였다.

한 것으로, 생성형 AI의 생산성 증가 효과가 뚜렷한 과업(task inside the frontier)과 그렇지 않은 과업(task outside the frontier)로 구성하였다. 즉, 과업1은 생성형 AI 활용에 따른 생산성 증가가 뚜렷한 과업인 반면, 과업2는 그렇지 않다. 즉, 과업1의 성격이 단순한 설명문의 정보 요약이라는 점에서 생성형 AI 활용이 즉각적으로 생성을 증가시킬 수 있으나, 구매층 특정 및 구매설득 이메일을 작성하는 과업2의 경우 창의성 및 정보의 재조합 등 고려해야 하는 조건들이 복잡하여 생성형 AI 활용이 해당 과업 수행을 반드시 도울 수 있다는 보장이 없다.

먼저 실험의 대조군은 분업이 존재한 상태에서 AI를 활용하지 않고 본 과업을 수행하였다. 대조군의 경우, 생성형 AI가 주어지지 않은 상태에서 먼저 약 90명의 실험참여자가 과업1을 수행한 뒤, 과업1의 결과물을 다시 다른 약 90명의 실험참여자에게 무작위로 배포한 뒤 과업2를 수행하였다. 따라서 마케팅 과업에서 분업이 실시되었고, 과업2를 수행하는 실험참가자는 원래 1,200자 분량의 설명문에 대한 정보는 없는 상태에서 제품의 설명문 요약만을 가지고 구매층 특정 및 구매설득 이메일을 작성하게 된다.

다음 실험군1의 경우 분업이 존재하는 상태에서 생성형 AI를 활용하여 본 과업을 수행하였다. 대조군처럼 과업1을 수행하는 집단과 과업2를 수행하는 집단이 나뉘어 있고, 과업2를 수행하는 집단은 과업1의 결과물이 무작위로 배포된 후 과업2를 수행한다. 하지만, 과업1과 과업2를 수행하는 모든 집단이 각 과업 수행 시 생성형 AI를 활용할 수 있는 환경이 제공되었다. 또한, 생성형 AI의 활용을 독려하기 위하여, 실험 화면에서 “본 과업을 수행하는 동안 생성형 AI(ChatGPT)가 제공됩니다. 생성형 AI를 적극 활용하여 본 과업을 수행하여 주십시오.”라는 문구를 추가하였다.

[그림 3-1] 실험에서 사용된 설명문

오르비스파(OrbiSpa)는 집에서도 스파의 편안함을 경험할 수 있도록 설계된 개인용 스마트 족욕기다. 제조사는 이 제품을 단순한 온수 기기가 아닌, 사용자의 일상 속 피로 회복과 심리적 안정을 돕는 '홈 스파 플랫폼'으로 정의한다. 빠르게 성장하는 웰니스 시장에서 오르비스파는 하루의 긴장을 풀고 스스로를 돌보는 시간을 제안하며, 기존 족욕기의 단순한 온열 기능을 넘어선 새로운 생활가전으로 주목받고 있다.

오르비스파는 최신 기술을 도입하여 제품의 가치를 극대화하였다. 내장된 센서가 사용자의 습관과 선호를 학습해 자동으로 온도와 시간을 조절하며, 별도의 조작 없이도 일정한 족욕 환경을 제공한다. 스마트폰 앱과 연동되어 예약 가동이나 원격 제어가 가능하고, 사용 기록을 통해 개인의 사용 패턴을 확인할 수도 있다. 본체는 재활용 합성수지 소재로 만들어져 내구성이 높고, 자동 순환 펌프가 물의 흐름을 일정하게 유지해 온도가 오래 지속된다. 전용 솔트 캡슐을 사용하면 미세한 기포와 음이온이 발생해 각질 제거와 피부 진정 효과를 주며, 사용 후에는 UV-C LED 살균 기능이 작동해 위생을 유지한다.

이 제품은 사용 과정에서 소비자 경험을 극대화할 수 있도록 설계되었다. 자동 온도 조절과 순환 기능 덕분에 물을 여러 번 교체할 필요가 없으며, 장시간 사용에도 일정한 온도를 유지해 관리가 간편하다. 살균 기능이 내장되어 별도의 세척 없이도 청결한 상태를 유지할 수 있고, 위생에 민감한 사용자에게 신뢰감을 준다. 재활용 합성수지와 내열 코팅 소재는 제품의 내구성을 높여 변색이나 손상 역시 최소화하였다.

하지만, 일반 족욕기보다 세 배 이상 높은 18만 9,000원의 가격은 진입 장벽으로 작용하며, 6리터 용량의 수조는 협소한 공간에서 사용과 보관을 어렵게 만든다. 전용 솔트 캡슐을 지속적으로 구입해야 하는 구조는 유지비 부담을 초래하고, 살균 과정에서 발생하는 소음과 전력 소모 역시 일부 사용자에게 불편함을 준다. 그럼에도 오르비스파는 단순한 가전을 넘어 개인의 회복과 돌봄을 위한 새로운 일상 도구로 평가된다. 기술적 안정성과 청결 관리, 그리고 사용 편의성을 결합한 설계는 '프리미엄 셀프케어'라는 새로운 생활양식을 제안한다. 가격과 실용성의 균형이라는 과제가 남아 있지만, 오르비스파가 상징하는 핵심 가치는 분명하다. 그것은 기술이 일상 속 자기관리의 방식을 바꾸고 있다는 사실이다.

자료: 저자작성

대조군과 실험군1 모두 분업이 존재하며 차이점은 생성형 AI의 활용 여부이다. 대조군과 실험군1의 참여자 모두 분업이라고 하는 상황을 인지하고 있는 상태에서 각 과업을 수행하였다. 즉, 과업1을 수행하는 집단은 자신의 과업 결과물이 다른 실험참여자에게 전달되어 구매층 특정 및 구매설득 이메일 작성에 활용된다는 사실을 인지한 상태에서 과업1을 수행하였으며, 과업2를 수행하는 참여자 역시, 자신이 현재 활용하고 있는 과업1의 결과물(요약문)이 어떤 실험참가자가 1,200자 분량의 제품설명문을 읽고 요약한 결과물이라는 것을 인지한 상태에서 구매층

특정 및 구매설득 이메일을 작성하였다.

다음으로 실험군2는 분업이 없는 상태에서 생성형 AI를 활용하지 않는 집단이다. 실험군2는 과업1과 과업2를 한 명의 실험참가자가 연속으로 수행하였고, 모든 과업에서는 생성형 AI가 주어지지 않았다. 참가자들은 실험 링크를 전달받고, 해당 링크에 접속하여 과업을 수행하였는데, 해당 링크 안에서는 생성형 AI 활용이 불가능하도록 설계하였다. 또한, 실험화면에서는 "본 과업을 수행하는 동안 다른 프로그램이나 다른 창을 여는 것은 엄격히 제한됩니다. 추후 다른 프로그램 사용 흔적이 발견될 경우, 추가 답례는 지급되지 않습니다."라는 문구를 추가하여 생성형 AI를 활용할 유인을 최소화하고자 했다. 실험군2의 참여자는 먼저 1,200자 분량의 설명문을 읽고 실험화면에서의 지시에 따라 요약한 뒤, 과업2를 수행할 때는 자신이 과업1에서 작성한 요약문에만 의존하여 구매층 특정 및 구매설득 이메일 작성을 수행하였다.

마지막으로 실험군3은 분업이 없는 상태에서 생성형 AI를 활용하는 집단이다. 실험군3 역시 실험군2와 같이 과업1과 과업2를 한 명의 참여자가 연속으로 수행하지만, 모든 과업을 수행할 때 실험화면에 ChatGPT를 활용할 수 있는 환경을 제공하고 ChatGPT 사용을 독려할 수 있는 문구를 실험군1과 같이 추가하였다. 실험군3 역시 과업2를 수행할 때는 제품설명문이 아닌, 자신이 요약한 자료만을 활용할 수 있도록 하였다.

실험군2와 실험군3의 경우 분업이 존재하지 않기 때문에 자신이 새로운 제품에 대한 정보를 직접 요약하고, 그 요약문을 바탕으로 구매층 특정 및 구매설득 이메일을 작성하게 된다. 따라서 분업이 존재할 경우 발생할 수 있는 정보의 손실을 최소화하게 된다. 분업이 있는 경우, 설명문 요약 과정에서 정보의 손실이 발생할 수 있고, 손실된 정보를 다시 복구할 수 있는 경로가 존재하지 않지만 실험군2와 실험군3의 경우, 과업2를 수행할 때 설명문을 다시 참조할 수는 없으나 자신이 직접 정보를 요약하였기에 정보손실이 존재할 가능성이 더 낮다.

〈표 3-1〉 실험설계 요약

구분	분업 존재			분업 없음
생성형 AI 비활용	대조군			실험군2
	과업1 (hteam1)	→	과업2 (hteam2)	과업1+과업2 (hindv)
생성형 AI 활용	실험군1			실험군3
	과업1 (aiteam1)	→	과업2 (aiteam2)	과업1+과업2 (aiindv)

자료: 저자작성

3. 검증 가능 가설

위와 같은 실험설계를 통해 다음과 같은 분석이 가능해진다. 먼저 대조군(hteam)과 실험군2(hindv)를 비교하여 분업 유무에 따른 생산성 변화를 살펴볼 수 있다. 대조군(hteam)과 실험군2(hindv)의 경우, 두 집단 모두 생성형 AI를 활용하지 않은 상태에서, 대조군(hteam)은 두 명의 다른 실험참여자가 과업1과 과업2를 수행하며 분업이 이루어지고, 실험군2(hindv)의 경우 한 명의 실험참여자가 과업1과 과업2를 모두 수행한다. 따라서 과업의 최종 결과물을 비교함으로써 분업의 비효율성 존재 여부를 검증할 수 있다.

가설 1. 대조군(hteam)과 실험군2(hindv)는 과업1 결과물의 품질에는 차이가 없지만, 과업2 결과물은 실험군2(hindv)의 품질이 대조군(hteam)보다 높다. 두 과업 모두 과업에 소요된 시간에는 큰 차이가 존재하지 않는다.

대조군(hteam)과 실험군2(hindv)의 차이는 분업의 유무만 존재하기 때문에 앞서 언급한 분업으로 인한 정보손실 또는 정보해석비용 등 분업의 부정적인 효과만 존재한다면, 대조군(hteam)과 실험군2(hindv)의 과업1 결과물에서는 차이가 존재하지 않지만, 분업의 부정적인 효과로 인해 대조군(hteam)의 과업2 결과물은 실험군2(hindv)의 과업2 결과물보다 낮은 품질을 가져야 한다. 또한 두 과업 모두 동일한 조건에서 생성형 AI의 활용 없이 수행되었기에 과업에 소요된 시간은 두 과업 모두 두 집단이 비슷해야 할 것이다.

다음으로 대조군(hteam)의 과업 결과물과 실험군3(aiindv)의 과업 결과물의 비교를 통해, 생성형 AI 활용 여부와 분업 유무가 과업결과물의 질과 소요시간에 미치는 영향을 살펴볼 수 있다. 이 경우, 생성형 AI 활용 여부와 분업 유무가 생산성에 미치는 영향이 뒤섞여 나타날 수 있으나 전체적인 비교가 가능하다.

가설 2-1. 실험군3(aiindv)의 과업1 결과물이 대조군(hteam)의 과업1 결과물보다 높은 품질을 가지며, 과업에 소요된 시간 역시 대조군보다 짧다.

가설 2-1은 실험군3(aiindv)과 대조군(hteam)의 과업1 결과물의 비교로, 생성형 AI가 과업1 수행에 즉각적인 영향을 준다는 가정하에서 도출되었다. 만약 생성형 AI가 과업1의 생산성을 높일 수 있다면, 생성형 AI를 활용하여 과업1을 수행한 실험군3(aiindv)이 대조군(hteam)보다 더 높은 품질의 요약문을 더 빠른 시간 내에 작성할 것이다.

가설 2-2. 실험군3(aiindv)의 과업2 결과물 품질이 대조군(hteam)의 과업2 결과물 품질보다 높으며, 실험군3(aiindv)이 과업 완료에 소요한 시간이 대조군(hteam)이 과업 완료에 소요한 시간보다 짧다.

가설 2-2는 분업이 생산성에 부정적인 영향이 있으며 과업2를 수행하는 데에는 생성형 AI가 큰 도움이 되지 않는다는 가정하에 작성되었다. 해당 가정에 따르면, 대조군의 경우 분업에 따른 정보손실 또는 해석비용 등으로 인한 부정적인 효과로 인해 과업2의 결과물 품질이 저해될 수 있으며 생성형 AI를 활용하지 못함으로써 발생하는 생산성 손실분은 존재하지 않을 것이다. 반면, 실험군3(aiindv)의 경우 분업으로 인한 부정적인 효과가 존재하지 않기에 과업2 결과물의 품질이 더 높을 것이다. 또한, 과업2 수행에서 생성형 AI를 활용함으로써 품질 측면에서의 도움은 없으나, 과업 수행 속도 측면에서는 생성형 AI 활용할 때 더 적은 시간이 소요될 것이다.

가설2에서 논의된 실험군3(aiindv)과 대조군(hteam)의 비교에서는 분업의 유무와 생성형 AI 활용 유무의 효과가 복합적으로 상호작용하고 있다. 이러한 복합적 상호작용을 살펴보기 위해서는 실험군2(hind)와 실험군3(aiindv)의 비교가 필

요하다. 실험군2(hindv)와 실험군3(aiindv)의 경우 모두 분업이 존재하지 않아, 분업으로 인한 영향의 간섭 없이 생성형 AI가 과업1과 과업2에 미치는 영향을 살펴볼 수 있다.

가설 3. 실험군2(hindv)의 과업1 결과물이 실험군3(aiindv)보다 낮은 품질을 가지며, 실험군2(hindv)의 과업2 결과물은 실험군3(aiindv)과 큰 차이가 없다. 각 과업에 소요된 시간은 실험군2(hindv)가 실험군3(aiindv)보다 더 길다.

가설 3은 생성형 AI가 과업1에는 긍정적인 영향을, 과업2에는 큰 영향을 미치지 않는다는 가정하에 도출되었다. 두 집단 모두 모든 참여자가 1,200자의 설명문을 직접 요약한 뒤, 요약한 정보를 바탕으로 구매설득 이메일을 작성하기에 정보손실 또는 정보해석 비용 등의 분업으로 인한 비효율성이 존재하지 않는다. 따라서 과업1의 결과물 품질이 곧바로 과업2의 결과물 품질과 반드시 양의 상관관계를 가질 필요는 없다. 다만, 생성형 AI를 활용할 경우 과업1에서는 긍정적인 영향이 존재하므, 과업1의 결과물 품질이 실험군3(aiindv)에서 더 높게 나타날 것이다. 하지만, 과업2의 경우 생성형 AI 활용이 제한적으로 생산성을 향상시키거나 그 효과가 없기 때문에 분업으로 인한 비효율성이 존재하지 않는 상황에서는 실험군3(aiindv)과 실험군2(hindv)의 과업2 결과물이 품질 측면에서 큰 차이를 나타내지 않을 것이다. 하지만 과업1과 과업2를 완료하는 데 소요된 시간은 생성형 AI를 활용하여 수행한 실험군3(aiindv)이 실험군2(hindv)보다 빠를 것이다. 이러한 비교를 통해서 AI 활용에 따른 과업1과 과업2에의 영향을 분업 효과의 간섭없이 살펴볼 수 있다.

다음으로 대조군(hteam)과 실험군1(aiteam)의 비교를 통해서는 생성형 AI가 각 과업에 갖는 효과가 분업과 어떻게 상호작용하는지를 살펴볼 수 있다. 두 집단 모두 분업이 존재하는 상황에서 각 과업 결과물의 품질을 살펴봄으로써 분업의 비효율성이 존재하는 상황에서 생성형 AI 활용이 각 과업에 미치는 영향을 살펴볼 수 있다.

가설 4. 대조군(hteam)의 과업1 결과물과 과업2 결과물의 품질 모두 실험군

1(aiteam)보다 낮다. 각 과업에 소요된 시간은 대조군(hteam)이 실험군1(aiteam)보다 더 길다.

가설 4는 생성형 AI 활용 여부가 분업의 비효율성에는 아무런 영향을 미치지 않는다는 가정과 생성형 AI의 활용이 과업1 결과물의 품질에는 긍정적인 영향을 주지만 과업2 결과물에는 긍정적인 영향이 없다는 가정하에 도출된 것이다. 먼저 생성형 AI가 과업1 수행에서 품질 개선에 영향을 미친다면, 과업1 결과물의 품질은 실험군1(aiteam)에서 더 높아야 한다. 더불어, 생성형 AI 활용 여부가 분업의 비효율성에 아무런 영향을 미치지 않고, 과업2 결과물이 과업1 결과물의 품질과 양의 상관관계가 있다면, 실험군1(aiteam)의 과업2 결과물의 품질이 대조군에 비하여 높아야 한다.

마지막으로 실험군1(aiteam)과 실험군3(aiindv)을 비교함으로써 생성형 AI가 분업의 효과에 미치는 영향을 살펴볼 수 있다. 실험군1(aiteam)과 실험군3(aiindv)의 경우 두 집단 모두 수행하는 모든 과업에서 생성형 AI를 활용하기 때문에 두 집단 간의 차이는 분업의 유무에 기인한다. 다만, 이 경우 가설 1과는 달리 두 집단 모두 생성형 AI를 활용하기 때문에 생성형 AI가 분업의 비효율성에 미치는 영향을 살펴볼 수 있다.

가설 5. 실험군1(aiteam)과 실험군3(aiindv)의 과업1 결과물의 품질에는 큰 차이가 존재하지 않지만, 과업2 결과물의 품질은 실험군3(aiindv)이 더 높다. 과업1과 과업2를 완료하는 데 소요된 시간은 두 집단 간 차이가 없다.

가설 5는 생성형 AI의 활용이 분업 효과에 직접적으로 영향을 미치지 않는다는 가정하에서 도출되었다. 생성형 AI의 활용이 분업의 비효율성 또는 부정적인 효과에 아무런 영향을 미치지 않는다면, 분업의 부정적인 효과와 상관없는 과업1의 경우, 실험군1(aiteam)과 실험군3(aiindv)에서 동일한 품질의 결과물이 관측되어야 할 것이다. 또한, 생성형 AI 활용이 분업의 비효율성에 아무런 영향을 미치지 않고 그대로 유지시킨다면, 과업2의 경우 분업이 없는 실험군3(aiindv)의 결과물 품질이 분업이 있는 실험군1(aiteam)에서 얻은 결과물의 품질보다 높아야 한다.

마지막으로 두 집단 모두 생성형 AI를 사용하여 모든 과업을 수행하였으므로, 두 과업에 소요된 시간은 모두 동일해야 한다.

제 2 절 실험 결과 분석

1. 실험 실사

본 연구의 실험은 모두 온라인으로 진행되었다. 제1절에서 살펴본 실험설계대로 각 집단별로 과업 수행에 대한 지시와 수행이 가능한 웹페이지를 구현하고, 해당 웹페이지의 링크를 실험참가자에게 전송하여 실험을 진행하였다. 설문조사 전문기관인 컨슈머인사이트에서 전국의 만 19세 이상 만 60세 미만의 모집단을 구성하고, 이들을 대상으로 실험 내용에 대해 설명하고 실험에 참가할 의사가 있음을 밝힌 사람들을 대상으로 총 6개의 실험링크를 배포하여 실험을 진행하였다. 총 602명이 실험에 참여하였고, 이 중 실험 과업과 관련없는 응답을 작성하거나 이메일 작성 과업을 완료하지 않은 참가자를 제외한 553명의 참가자가 실험을 완료하였다. 모든 실험참가자들은 실험을 완료하면 5,000원의 기본 보상과 자신의 과업 결과물의 평가에 따라 추가 보상이 존재함을 인지한 상태에서 실험에 참여하였다. 또한, 과업 결과물의 평가는 과업 결과물의 품질과 과업 완료에 소요되는 시간에 따라 이루어진다는 사실 역시 공지하여 이를 모두 인지한 상태에서 실험에 참여하였다. 대조군(hteam)과 실험군2(aiteam)의 경우 과업1과 과업2를 수행한 집단이 다르고, 실험참가자 1명당 하나의 과업을 수행하였기 때문에 추가 보상은 과업결과물의 평가에 따라 4,000원, 9,000원, 14,000원이 차등적으로 지급되었다. 실험군1(hindv)과 실험군3(aiindv)의 경우, 과업1을 수행한 뒤에 과업2를 수행하여 실험참가자 1인당 2개의 과업을 수행하였는데, 각 과업별로 평가에 따라 4,000원, 9,000원, 14,000원이 차등적으로 지급되어 과업1과 과업2에서 모두 좋은 평가를 얻은 경우 최대 28,000원이 추가 지급되었다. 즉, 각 과업별로 평가와 지급을 분리하여 분업이 존재하지 않는 경우와 분업이 존재하는 경우를 비교할 때, 실험참가자의 과업 수행 유인이 동일하게 유지될 수 있도록 추가 보상을 설정하였다.

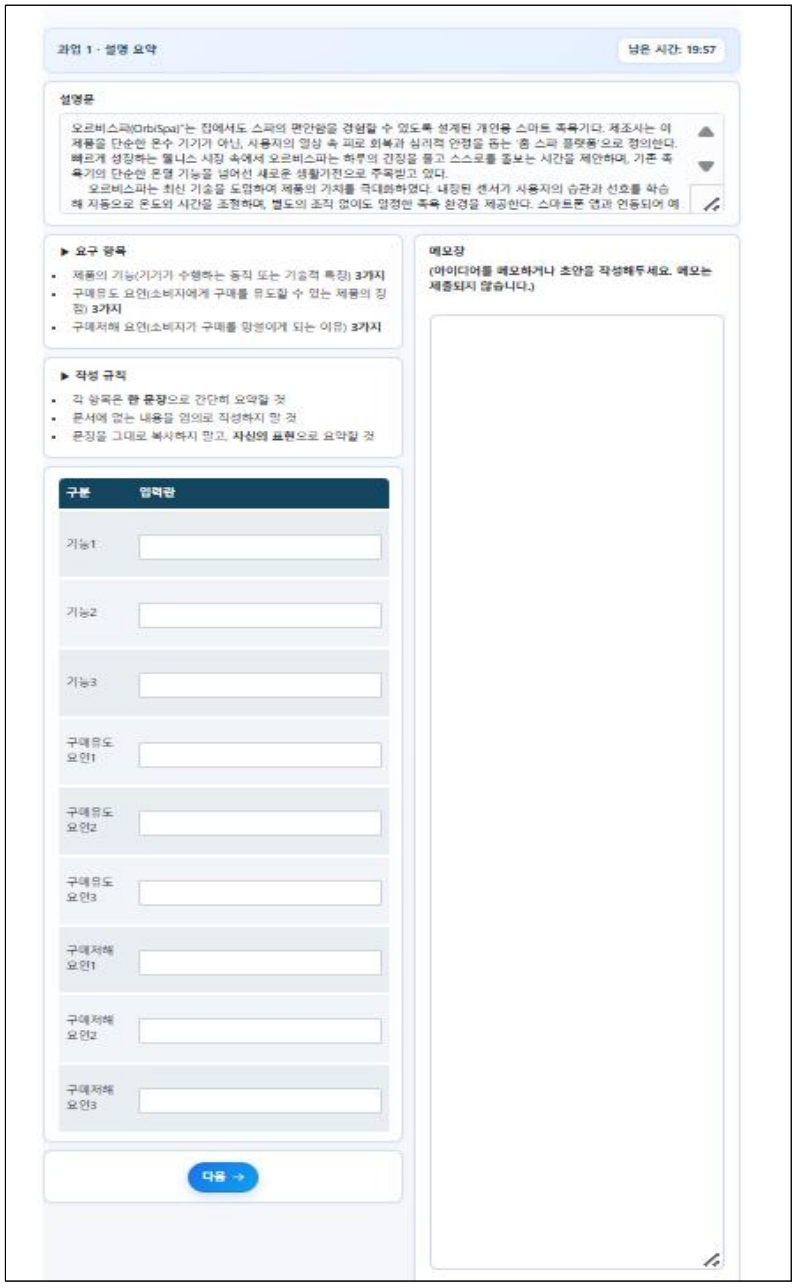
실험은 총 6개의 링크를 무작위로 실험 참가에 동의한 인원을 대상으로 배포하여 진행되었다. 대조군(hteam)과 실험군1(aiteam)의 경우 과업1과 과업2를 수행하는 실험참가자가 다르고 분업이 존재하기 때문에, 각 집단별로 과업1과 과업2의 실험링크가 필요하였다. 또한, 해당 두 집단의 실험은 순차적으로 진행되어, 대조군과 실험군1에서 각각 과업1을 수행하는 실험을 진행한 뒤에, 과업결과물을 정리하여 각 과업결과물을 과업2를 수행하는 실험에 참가하는 인원을 대상으로 무작위로 매칭한 뒤 과업2 실험을 수행하였다. 실험군1(hindv)과 실험군3(aiindv)의 경우 분업이 존재하지 않기 때문에, 각각 하나의 실험링크를 무작위로 배포하여 실험을 진행하였다.

생성형 AI를 활용하는 집단인 실험군1(aiteam)과 실험군3(aiindv)의 경우 실험화면에 직접 ChatGPT를 사용할 수 있는 창을 구현하여, 과업을 수행하며 직접 활용할 수 있는 환경을 제공하였고, 생성형 AI를 활용하지 않는 집단의 경우에는 ChatGPT 창과 동일한 모양과 형태를 한 메모장을 제공하여 생성형 AI 활용을 제외한 모든 과업 수행 환경을 동일하게 하고자 하였다.

[그림 3-2] 실험 화면 스크린샷: 생성형 AI 활용

자료: 저자작성

[그림 3-3] 실험 화면 스크린샷: 생성형 AI 비활용



자료: 저자작성

2. 과업 결과물의 평가

모든 집단의 과업결과물의 품질은 모든 실험이 완료된 이후에 ChatGPT를 활용하여 평가하였다. 모든 과업 결과물이 문자로 이루어져 있기에, LLM을 활용한 실제의미 기반(semantic scoring)의 채점 방식을 활용하여 평가를 진행하였다. 실제의미 기반의 채점 방식은 기존의 특정 단어가 포함 여부를 바탕으로 평가하는 채점방식에서 벗어나, 문장 전체가 전달하는 의미를 LLM이 직접 이해하여 점수를 산정하는 방식으로 실제 인간이 문장을 읽고 평가하는 방식과 매우 흡사한 결과를 도출할 수 있다. 과업1의 경우, 정확성, 완성도, 표현력의 세 항목을 기준으로 아래와 같은 채점기준(Rubric)을 활용하여 과업결과물의 평가가 이루어졌다.

[그림 3-4] 과업1 결과물의 품질 평가를 위한 채점기준

① 정확성
- 1점: 주요 기능·구매유도요인·구매저해요인의 의미가 왜곡되거나 누락 다수
- 3점: 핵심 내용은 포함했으나 기술적 세부나 용어 사용이 부정확함
- 5점: 핵심 내용은 대체로 정확하나 일부 단순화 또는 세부 불일치 존재
- 7점: 원문 의미를 정확히 이해·전달하며, 세부 기술까지 정밀하게 반영함

② 완성도
- 1점: 6개 이하 작성 또는 한 범주(기능/효용/제약)에 편중됨
- 3점: 7~8개 작성, 일부 범주 불균형 또는 내용 빈약
- 5점: 9개 항목 모두 작성하였으나 세부 내용의 깊이에 차이 있음
- 7점: 9개 항목 모두 충실히 작성하고 세 영역의 정보가 균형 있게 포함됨

③ 표현력
- 1점: 원문 문장을 그대로 복사하거나 단순 나열(정보 재구성 없음)
- 3점: 일부 재구성했으나 문장 간 연결과 논리 흐름이 약함
- 5점: 문장 단위로 요약·재구성하며 전체 의미 흐름을 유지함
- 7점: 정보를 논리적으로 재배열·통합하여 간결하고 구조적으로 요약함

자료: 저자작성

과업1 결과물의 품질은 위 세 항목 점수의 단순평균으로 측정하였다.

과업2의 결과물 역시 실제의미 기반 방식으로 평가하였는데, 과업2 결과물의 품질 평가는 특정한 구매층의 구체성, 이메일의 논리성, 명료성, 설득력, 톤/적합

성의 다섯 항목으로 이루졌으며, 각 항목별로 아래와 같은 채점기준이 사용되었다.

[그림 3-5] 과업2 결과물의 품질 평가를 위한 채점기준

① 세그먼트 구체성
- 1점: 미작성/불완전
- 3점: 일부만 제시, 이유 없음
- 5점: 연령·성별·직업 모두 작성, 이유 모호
- 7점: 모두 구체적, 선택 이유가 제품 특징과 논리적으로 연결

② 논리성(제품 특징 ↔ 고객 니즈 연결)
- 1점: 연결 없음
- 3점: 일부 연결 시도있지만 불명확
- 5점: 대체로 연결되나 설득력 약함
- 7점: 니즈와 장점이 명확히 대응, 논리적·강력

③ 명료성
- 1점: 장황/구조 불분명
- 3점: 구조 있으나 반복·모호함 많음
- 5점: 대체로 명확하나 일부 장황
- 7점: 간결·깔끔, 가독성 뛰어남

④ 설득력(장점 강조 + 구매저해요인 언급 및 상쇄 방안 제시 여부)
- 1점: 장점 거의 없음, 구매저해요인 언급/대응 없음
- 3점: 장점 있으나 구매저해요인 상쇄 방안 없음
- 5점: 장점+구매저해요인 언급 있으나 상쇄 방안 약함
- 7점: 장점이 강하게 드러나고 구매저해요인 언급하며 상쇄 방안까지 설득력 있게 제시

⑤ 톤/적합성(세그먼트와 메시지 일치)
- 1점: 타깃에 어울리지 않는 어투·표현
- 3점: 일부 맞춤 요소 있으나 일반적
- 5점: 대체로 적합하나 호소력 약함
- 7점: 세그먼트 특성에 맞는 어투·표현, 일관성 유지

자료: 저자작성

과업2 결과물의 품질은 위 다섯 항목 평가점수의 단순 평균으로 측정되었다. 모든 과업 결과물은 동일한 기준으로 평가되었으며 채점에는 생성형 AI 활용 유무나 분업 유무에 대한 정보가 활용되지 않았다.

다음으로 각 과업에 소요된 시간은 실험참가자가 과업을 실시하는 웹페이지를 보기 시작한 시간과 해당 페이지에서 벗어나는 시간의 차이를 기록하고 두 시간

의 차이를 활용하여 측정하였다. 과업1의 경우 설명문 요약 페이지에 접근한 시간부터 설명문 요약 페이지를 벗어나는 시간이 과업1의 총소요시간이며, 과업2의 경우, 구매층 특정 페이지에 접근한 시간부터 이메일 과업 페이지를 벗어난 시간이 총소요시간이다.

3. 실험 분석 결과

가설 검증에 앞서 실험에 참여한 응답자의 인구 통계 특성과 AI 활용 특성을 살펴보았다. 응답자 558명 중 남성과 여성은 각 50.0%이며, 연령은 20대부터 50대에 걸쳐 고르게 분포하였다. 직업은 사무직이 71.1%, 교육 수준은 대학 재학 이상이 94.4%로 높은 비중을 차지하였다. AI 활용 측면에서 AI 도구를 주 2~3회 활용한다고 응답한 비율이 전체의 69.9%로 나타났으며, 그중 매일 활용한다는 응답도 20.8%로 조사되었다. AI 도구 숙련도 측면에서 잘 활용하지 못한다고 응답한 비율은 4.8%에 불과하여 응답자의 95.2%가 보통 수준 이상으로 활용한다고 인식하는 것으로 나타났다.

〈표 3-2〉 응답자 일반 특성과 AI 활용 특성

구분			사례 수(명)	비율(%)
일반 특성	성별	남	279	50.0
		여	279	50.0
	연령	20대	110	19.7
		30대	156	28.0
		40대	156	28.0
		50대	136	24.4
	직업	사무직	397	71.1
		비사무직	65	11.6
		무직	96	17.2
	교육수준	고졸 이하	31	5.6
		대재 이상	527	94.4

구분			사례 수(명)	비율(%)
AI 활용 특성	AI 사용 빈도	매일	116	20.8
		주 4~5회	115	20.6
		주 2~3회	159	28.5
		주 1회	88	15.8
		월 3회	29	5.2
		월 2회 이하	51	9.1
	AI 사용 숙련도	잘 활용하지 못함	27	4.8
		보통	315	56.4
		잘 활용함	170	30.5
		매우 잘 활용함	46	8.2

자료: 저자작성

모든 실험참가자가 수행한 과업 결과물의 품질 및 소요시간에 대한 데이터를 확보한 뒤, 앞서 살펴본 검증 가능한 가설들을 검증하기 위한 통계적 검증을 실시하였다. 본 연구에서는 단순한 두 집단의 비교가 주를 이루기에, 두 집단의 과업 결과물의 품질, 즉 ChatGPT의 항목별 평가점수의 평균, 비교와 과업에 소요된 시간을 비교하기 위하여 간단한집단 크기가 상이한 경우의 t-검정(Welch's Two-Sample t-test)을 활용하여 분석을 진행하였다. 추가적으로 두 집단 간의 과업결과물 품질 분포 비교를 위해 K-S 검정(Kolmogorov-Smirnov test)과 윌콕슨 순위합 검정(Wilcoxon Rank-Sum Test)을 활용하였다.

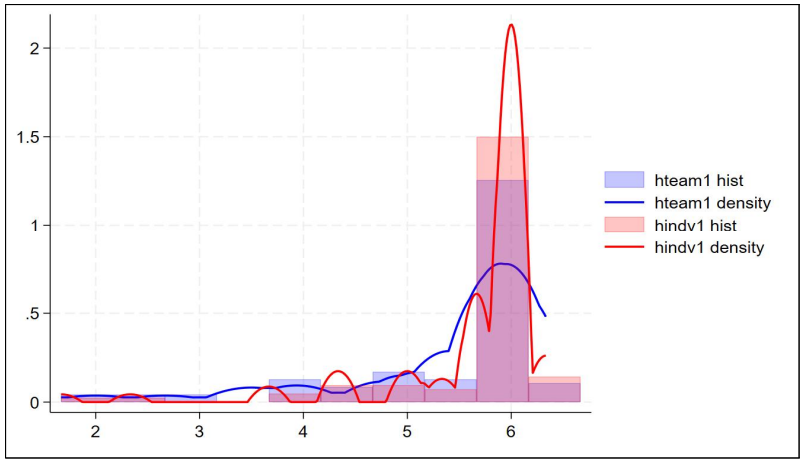
가설1의 검증을 위하여 대조군(hteam)과 실험군2(hindv)의 각 과업 결과물의 품질과 소요시간을 비교하였다. 먼저 대조군(hteam)과 실험군2(hindv)의 과업1 결과물의 품질을 비교한 t-검정 결과 실험군2(hindv)의 과업1 결과물의 품질이 5% 유의수준에서 더 높은 것으로 나타났다. [그림 3-6]에서 볼 수 있듯이, 두 집단의 과업결과물 품질 분포의 형태 비교를 위한 KS검정에서는 두 집단의 분포의 형태가 다르지 않으며, 순위합 검정 결과 실험군2(hindv)의 분포가 대조군(hteam)에 비하여 10% 유의수준에서 더 오른쪽으로 이동한 경향이 있는 것으로 나타났다.

〈표 3-3〉 과업1의 품질: 대조군(hteam)과 실험군2(hindv)

집단	관측 수	평균(점)	표준편차
대조군 (hteam)	94	5.386525	1.055569
실험군2 (hindv)	84	5.666667	.7948967
t-검정(p-value)	0.0457		
K-S검정(p-value)	0.345		
순위합검정(p-value)	0.0509		

자료: 저자작성

[그림 3-6] 과업1 결과물의 품질: 대조군(hteam) vs. 실험군2(hindv)



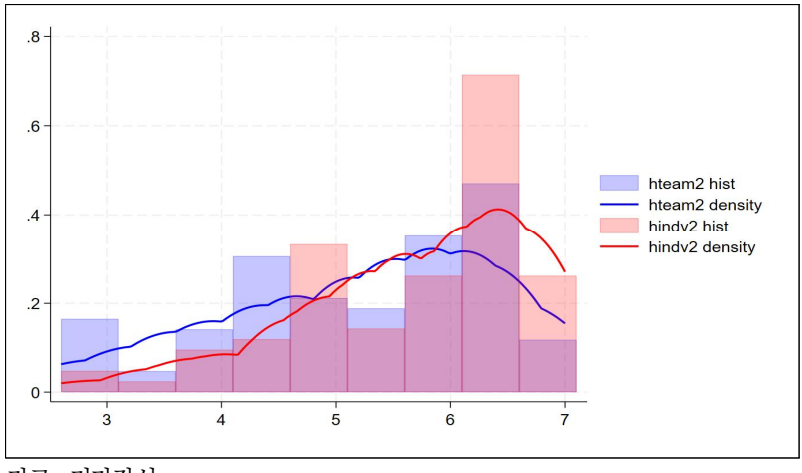
다음으로 두 집단 간의 과업2 결과물을 비교한 t-검정 결과, 실험군2(hindv)의 과업 결과물 품질이 대조군(hindv)에 비하여 1% 유의수준에서 더 높은 것으로 나타났다. 품질 분포의 비교에서도 K-S 검정 결과 두 집단의 분포 형태가 10% 유의수준에서 다르며, 순위합 검정 결과 실험군2(hindv)의 분포가 대조군(hindv)보다 1% 유의수준에서 더 오른쪽으로 이동한 경향이 있는 것으로 나타났다.

〈표 3-4〉 과업2의 품질: 대조군(hteam)과 실험군2(hindv)

집단	관측 수	평균(점)	표준편차
대조군 (hteam)	85	5.230588	1.19742
실험군2 (hindv)	84	5.738095	1.034339
t-검정(p-value)	0.0036		
K-S검정(p-value)	0.083		
순위합검정(p-value)	0.0046		

자료: 저자작성

[그림 3-7] 과업2 결과물의 품질: 대조군(hteam) vs. 실험군2(hindv)



자료: 저자작성

가설 1에 따르면, 분업에 의한 비효율성만이 존재하기 때문에 과업1에서는 두 집단 간 차이가 없고, 과업2에서만 두 집단 간 결과물 품질에 차이가 유의미해야 하지만, 실제 실험 결과 분업이 없는 경우에 과업1의 품질에서도 유의미하게 차이가 발생하였다. 즉, 분업이 없는 경우에 과업1의 품질이 통계적으로 유의하게 높은 것인데, 이는 특정 행태적 편향(behaviroal bias)으로 설명될 수 있다. 분업이 존재하지 않는 경우에는 자신이 과업1의 결과물을 활용하여 과업2를 수행해야 하

는 것을 인지하고 있기에, 과업1을 더 성실히 수행하는 반면, 분업이 존재할 경우 자신의 과업1 결과물이 다른 누군가가 수행할 과업2의 기초자료가 된다는 사실을 인지하고 있으나, 이들에게는 과업1의 종료가 곧 자신의 모든 과업의 종료이기에 분업이 존재하지 않는 경우보다 더 낮은 노력으로 덜 성실히 과업1을 수행할 수 있다.

흥미로운 점은 대조군(hteam)과 실험군2(hindv)에서 과업1에 대한 추가보상 체계가 동일함에도 불구하고, 자신이 과업1의 결과물을 추후 사용할 것인지 타인이 사용할 것인지에 따라 과업1 결과물에 차이가 있다는 것이다. 추후 논의하겠지만, 이러한 행태적 편향은 실험군1(aiteam)과 실험군3(aiindv)에서도 동일하게 관측되어 대조군(hteam)과 실험군2(hindv)의 비교에서만 나타나는 특징적인 현상은 아니다.

과업2의 결과물 품질은 가설 1에서 예상한 대로 분업으로 인한 비효율성으로 인해 실험군2(hindv)에서 더 높게 나타났다. 기존에 존재하던 정보손실 또는 정보해석 비용에 더하여, 앞서 살펴본 행태적 편향에 따른 과업1 결과물의 차이까지 반영되어, 대조군(hteam)과 실험군2(hindv)의 과업2 결과물 품질의 차이를 더 악화시켰을 가능성이 있다. 즉, 기존 분업의 부정적 효과인 정보손실과 해석비용 등에 더하여 과업1 수행 시 분업에 의한 행태적 편향이 두 집단 간 과업2 품질의 차이를 더 크게 만들었을 수 있다.

마지막으로 과업 완료에 소요된 시간은 과업1과 과업2 모두 대조군(hteam)과 실험군2(hindv) 간에 차이가 없는 것으로 나타났다.

결과 1 (분업으로 인한 비효율성). 생성형 AI의 활용이 없는 경우, 분업으로 인한 비효율성이 존재한다. 분업하에서 자신의 과업이 최종 과업이 아님에서 기인하는 행태적 편향으로 인해 대조군(hteam)의 과업1 결과물의 품질이 실험군2(hindv)보다 낮다. 또한 분업으로 인한 정보손실 및 정보해석 비용 등으로 인해 대조군(hteam)의 과업2 결과물의 품질이 실험군(hindv)보다 낮다. 두 과업을 완료하는 데 소요된 시간에는 두 집단 간 유의미한 차이가 존재하지 않는다.

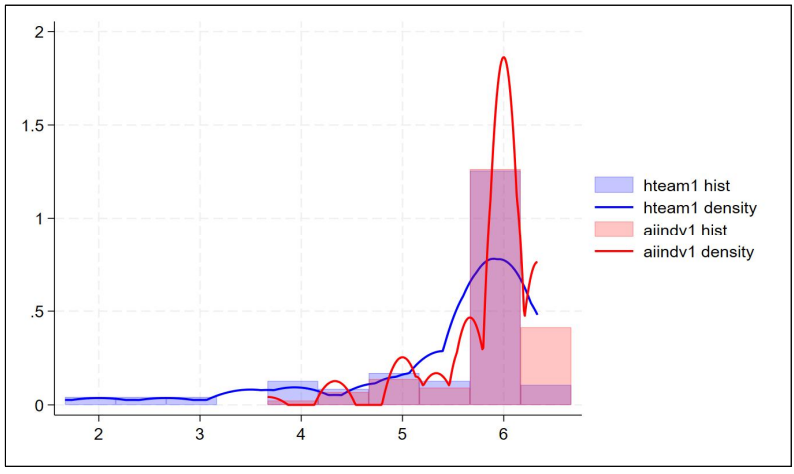
가설 2-1과 2-2는 분업 유무와 생성형 AI 활용 여부가 생산성에 미치는 복합적 영향에 관한 것으로, 대조군(hteam)과 실험군3(aiindv)을 비교할 필요가 있다. 가설 2-1에 따르면 대조군(hteam)의 과업1 결과물의 품질이 실험군3(aiindv)보다 낮고 과업1에 소요된 시간 역시 실험군3(aiindv)에서 더 적어야 한다. 두 집단의 과업1 품질을 비교한 t-검정 결과, 예상대로 실험군3(aiindv)의 품질이 1% 유의수준에서 더 높았다. 이러한 차이는 생성형 AI의 활용으로 인한 과업1 수행에서의 생산성 향상과, 앞서 살펴본 분업에 의한 과업1 수행에 있어서의 행태적 편향이 반영된 결과로 해석할 수 있다. 또한 K-S 검정 결과, 과업결과물 품질 분포 형태에는 두 집단간 차이가 유의미하지 않았으나, 순위합 검정결과에서는 실험군3(aiindv)의 품질 분포가 1% 유의수준에서 더 오른쪽으로 이동한 경향이 있는 것으로 나타나, t-검정과 유사한 결과가 나타났다. 과업에 소요된 시간 역시 가설 2의 예상대로 실험군3(aiindv)은 약 8분, 대조군(hteam)은 약 7분으로 실험군3(aiindv)에서 더 짧게 나타났으나 통계적으로 유의하지 않았다.

〈표 3-5〉 과업1의 품질: 대조군(hteam)과 실험군3(aiindv)

집단	관측 수	평균(점)	표준편차
대조군 (hteam)	94	5.386525	1.055569
실험군3 (aiindv)	87	5.842912	.5129816
t-검정(p-value)	0.0003		
K-S검정(p-value)	0.105		
순위합검정(p-value)	0.0003		

자료: 저자작성

[그림 3-8] 과업1 결과물의 품질: 대조군(hteam) vs. 실험군3(aiindv)



자료: 저자작성

〈표 3-6〉 과업1의 소요시간: 대조군(hteam)과 실험군3(aiindv)

집단	관측 수	평균(분)	표준편차
대조군 (hteam)	94	8.151722	5.210929
실험군3 (aiindv)	87	7.369988	4.408733
t-검정(p-value)	0.2762		

자료: 저자작성

다음으로 가설 2-2는, 생성형 AI 활용에 의한 긍정적인 효과는 과업2 수행에서 잘 나타나지 않으나, 분업의 부정적인 효과로 인해 실험군3(aiindv)의 과업2 결과물의 품질이 대조군(hteam)보다 높을 것이라고 예상하고 있다. 즉, 분업에서 오는 부정적인 효과가 생성형 AI의 도입에 따른 분업의 소멸로 상쇄되어 과업2의 품질이 향상될 것이라는 것이다. 또한 생성형 AI의 활용이 품질에는 긍정적인 영향을 미치지 않지만, 과업2 수행에 소요된 시간은 어느 정도 단축시키는 데 일조할 수 있기에 실험군3(aiindv)에서의 과업 완료에 소요된 시간이 더 적을 것이라

고 예상하였다.

실험군3(aiindv)과 대조군(hteam)의 과업2 결과물의 품질과 소요된 시간을 비교한 t-검정 결과, 과업2 결과물의 품질은 1% 유의수준에서 실험군3(aiindv)이 더 높고, 소요시간 역시 5% 유의수준에서 실험군3(aiindv)이 더 짧은 것으로 나타나, 가설 2-2를 뒷받침하는 결과를 얻었다. 과업결과물의 품질 분포 비교 결과, K-S 검정에서는 두 집단의 분포가 10% 유의수준에서 상이하며, 순위합 검정 결과 1% 유의수준에서 실험군3(aiindv)의 분포가 더 오른쪽으로 이동한 경향이 있는 것으로 나타났다.

〈표 3-7〉 과업2의 품질: 대조군(hteam)과 실험군3(aiindv)

집단	관측 수	평균(점)	표준편차
대조군 (hteam)	85	5.230588	1.19742
실험군3 (aiindv)	87	5.735632	1.024125
t-검정(p-value)	0.0034		
K-S검정(p-value)	0.064		
순위합검정(p-value)	0.0046		

자료: 저자작성

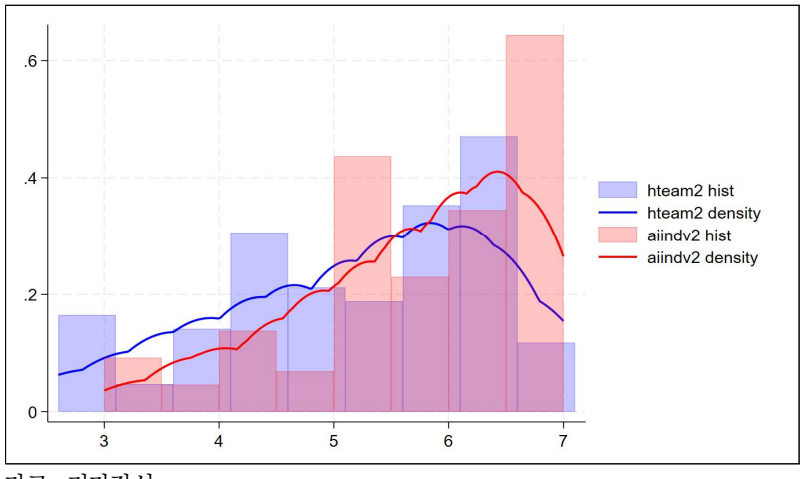
〈표 3-8〉 과업2의 소요시간: 대조군(hteam)과 실험군3(aiindv)

집단	관측 수	평균(분)	표준편차
대조군 (hteam)	85	16.35485	7.615256
실험군3 (aiindv)	87	13.64013	7.846716
t-검정(p-value)	0.0225		

자료: 저자작성

결과 2 (생성형 AI 활용과 분업의 소멸). 생성형 AI를 활용하고 분업이 없는 실험군3(aiindv)의 생산성이 대조군(hteam)보다 높다. 모든 과업 결과물의 품질은 실험군3(ainndv)이 대조군에 비하여 높으며, 과업2 완료에 소요된 시간 역시 더 적다. 실험군3(aiindv)의 과업1 결과물의 품질이 대조군(hteam)보다 높은 것은 분업의 부재로 대조군(hteam)에서 관측되는 과업1 수행에서의 행태적 편향이 존재하지 않고, 생성형 AI 활용으로 인한 생산성 향상에 기인한다. 실험군3(aiindv)의 과업2 결과물의 품질이 대조군(hteam)보다 높은 것은 분업의 소멸로 인한 분업의 부정적 영향이 제거된 것과 생성형 AI 활용으로 인한 생산성 향상에 기인한다.

[그림 3-9] 과업2 결과물의 품질: 대조군(hteam) vs. 실험군3(aiindv)



자료: 저자작성

가설 2에서 살펴본 생성형 AI의 활용과 분업의 소멸로 인한 생산성 향상효과는 앞서도 논의했듯이, 생성형 AI 활용 여부와 분업 유무가 서로 뒤얽혀 복잡하게 상호작용하며 나타난 결과로 해석에 유의해야 한다. 이처럼 복잡한 효과를 풀어서 명확히 살펴보기 위하여, 가설 3은 분업 측면에서의 간섭 없이 생성형 AI 활용에 대한 효과만을 살펴보기 위하여 실험군2(hindv)와 실험군3(aiindv)을 비교하였다. 즉, 두 집단 모두 분업이 존재하지 않는 상황에서, 생성형 AI의 활용이 과업1

결과물의 품질을 높이는 반면, 과업2 결과물의 품질에는 제한적이거나 미미한 영향을 미친다는 가정하에서, 실험군3(aiindv)의 과업1 결과물 품질은 실험군2(hindv)에 비하여 높고, 과업2 결과물의 품질에는 두 집단 간 큰 차이는 없을 것으로 예상하고 있다. 다만, 과업 완료 소요 시간의 경우 두 과업 모두에서 실험군3(aiindv)이 실험군2(hindv)보다 짧다고 예상하고 있다. 먼저 실험군2(hindv)와 실험군3(aiindv)의 과업1 결과물의 품질과 과업 완료에 소요된 시간을 비교한 t-검정 결과, 품질의 경우 10% 유의수준에서 실험군3(aiindv)이 더 높고, 과업 완료에 소요된 시간의 경우 5% 유의수준에서 실험군3(aiidnv)이 더 짧은 것으로 나타나 가설 3을 뒷받침하는 결과를 얻었다. 과업 1의 품질 분포 비교를 위한 K-S 검정 결과 두 집단의 품질 분포 형태는 상이하지 않으며, 순위합 검정 결과 10% 유의수준에서 실험군3(aiindv)의 품질 분포가 더 오른쪽으로 이동하여 실험군2(hindv) 분포보다 더 높은 값을 갖는 경향이 있는 것으로 나타났다.

〈표 3-9〉 과업1의 품질: 실험군2(hindv)와 실험군3(aiindv)

집단	관측 수	평균(점)	표준편차
실험군2 (hindv)	84	5.666667	.7948967
실험군3 (aiindv)	87	5.842912	.5129816
t-검정(p-value)	0.0883		
K-S검정(p-value)	0.413		
순위합검정(p-value)	0.0702		

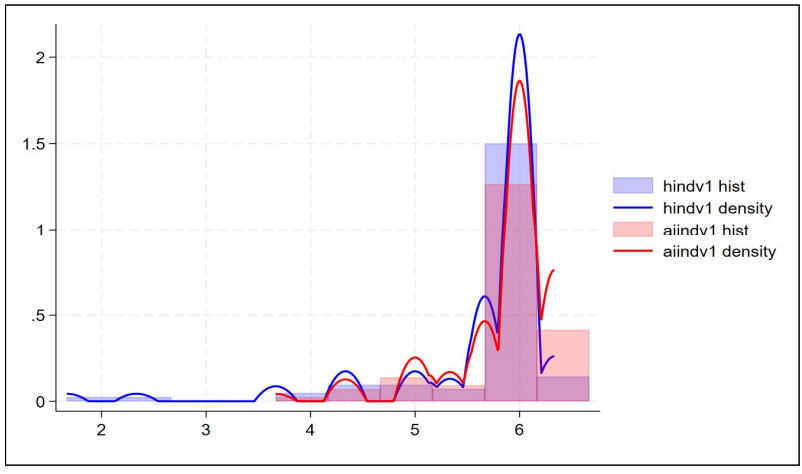
자료: 저자작성

〈표 3-10〉 과업1의 소요시간: 실험군2(hindv)와 실험군3(aiindv)

집단	관측 수	평균(분)	표준편차
실험군2 (hindv)	84	9.126043	5.283031
실험군3 (aiindv)	87	7.369988	4.408733
t-검정(p-value)	0.0197		

자료: 저자작성

[그림 3-10] 과업1 결과물의 품질: 실험군2(hindv) vs. 실험군3(aiindv)



자료: 저자작성

다음으로 실험군3(aiindv)과 실험군2(hindv)의 과업2 결과물을 비교를 위한 t-검정 결과, 품질에서는 큰 차이가 없었고, 소요시간은 실험군3(aiindv)이 실험군2(hindv)보다 평균적으로 2분 정도 빨리 과업을 완료하였으나 통계적으로 유의하지는 않았다. 두 집단간 품질 분포 비교를 위한 K-S 검정과 순위합 검증 결과 두 집단 간 결과물 품질 분포의 형태와 위치가 통계적으로 유의미하게 다르지 않은 것으로 나타나 t-검정 결과와 비슷한 결과를 확인할 수 있었다.

〈표 3-11〉 과업2의 품질: 실험군2(hindv)와 실험군3(aiindv)

집단	관측 수	평균(점)	표준편차
실험군2 (hindv)	84	5.738095	1.034339
실험군3 (aiindv)	87	5.735632	1.024125
t-검정(p-value)	0.9875		
K-S검정(p-value)	1.0000		
순위합검정(p-value)	0.9795		

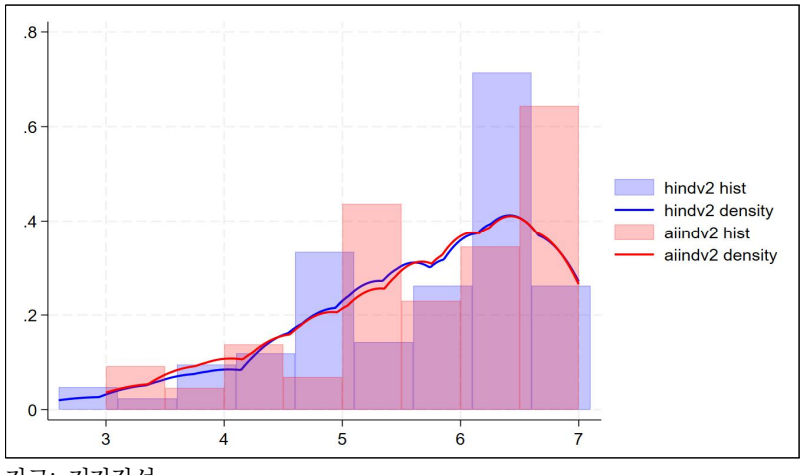
자료: 저자작성

〈표 3-12〉 과업2의 소요 시간: 실험군2(hindv)와 실험군3(aiindv)

집단	관측 수	평균(분)	표준편차
실험군2 (hindv)	84	15.18806	6.908465
실험군3 (aiindv)	87	13.64013	7.846716
t-검정(p-value)	0.1724		

자료: 저자작성

[그림 3-11] 과업2 결과물의 품질: 실험군2(hindv) vs. 실험군3(aiindv)



자료: 저자작성

위의 결과를 종합해 볼 때, 분업이 존재하지 않을 경우 생성형 AI의 활용은 과업1 결과물의 품질에는 긍정적인 영향을, 과업2 결과물의 품질에는 영향을 미치지 않는다는 가정이 틀리지 않았음을 알 수 있다.

결과 3 (생성형 AI 활용이 과업1과 과업2에 미치는 영향). 실험군3(aiindv)의 과업1 결과물의 품질은 실험군2(hindv)보다 높으나, 과업2 결과물의 품질에서는 두 집단 간에 큰 차이가 없다. 즉, 분업이 존재하지 않는 경우, 생성형 AI의 활용은 비교적 단순한 과업(과업1)의 생산성은 높이나 창의성 및 정보 재조합이 요구되는 복잡한 과업(과업2)의 생산성은 높이지 않는다.

다음으로 가설 4는 분업이 존재하는 경우, 생성형 AI의 활용이 각 과업에 미치는 영향에 대한 가설로, 생성형 AI 활용이 분업의 부정적인 효과에 전혀 영향을 미치지 않으며, 과업1에는 긍정적인 효과를, 과업2에는 제한적이거나 영향을 미치지 않는다는 가정하에서 도출되었다. 이미 두 번째 가정, 즉 생성형 AI의 활용이 과업1에는 긍정적인 영향을, 과업2에는 영향을 미치지 않음이 결과 3을 통해 검증되었기 때문에, 만약 검증 결과 가설 4와 같은 결과가 도출된다면 생성형 AI의 활용이 분업의 효과에는 영향을 미치지 않는다는 가정을 뒷받침하게 되고, 가설 4와 다른 결과가 도출된다면 생성형 AI 활용이 분업의 효과에 영향을 미치게 된다는 것을 뜻한다. 먼저 대조군(hteam)과 실험군1(aiteam)의 과업1 결과물을 비교한 t-검정 결과, 과업1 결과물의 품질과 과업 완료에 소요된 시간에서 유의미한 차이가 발견되지 않았다. 두 집단의 품질 분포 비교에서는, K-S 검정 결과 두 집단의 분포 형태 차이가 유의미하지 않으나, 순위합검정에서는 5% 유의수준에서 실험군1(aiteam)의 과업 1 결과물 품질 분포가 대조군(hteam) 보다 더 오른쪽에 위치하는 경향이 있는 것으로 나타났다.[10)]

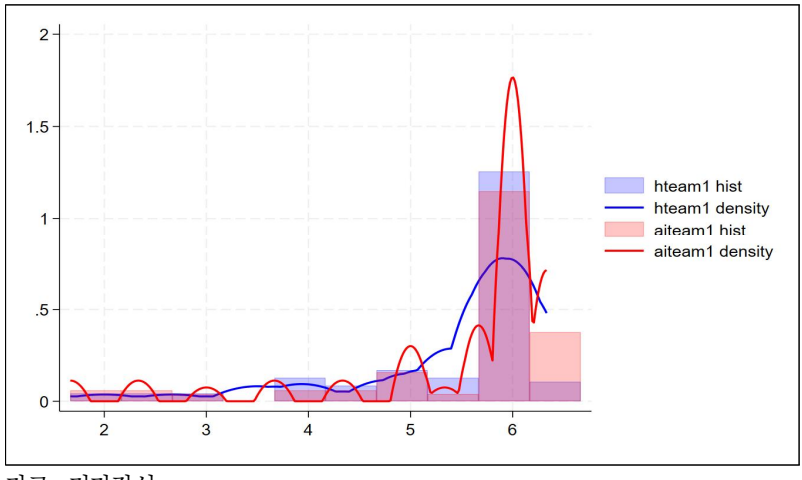
10) 이는 t-검정 결과와 반대되는 것으로, [그림 3 - 12]에서 볼 수 있듯이 분포 상 실험군1(aiteam)의 분포가 조금 더 오른 쪽에 위치하는 경향이 있으나, 실험군1(aiteam)의 낮은 품질의 과업 결과물이 대조군 보다 더 많아 평균적으로는 큰 차이가 나타나지 않는 것으로 보인다.

〈표 3-13〉 과업1의 품질: 대조군(hteam)과 실험군1(aiteam)

집단	관측 수	평균(점)	표준편차
대조군 (hteam)	94	5.386525	1.055569
실험군1 (aiteam)	101	5.518152	1.131837
t-검정(p-value)	0.4018		
K-S검정(p-value)	0.338		
순위합검정(p-value)	0.0255		

자료: 저자작성

[그림 3-12] 과업1 결과물의 품질: 대조군(hteam) vs. 실험군1(aiteam)



자료: 저자작성

〈표 3-14〉 과업1의 소요 시간: 대조군(hteam)과 실험군1(aiteam)

집단	관측 수	평균(분)	표준편차
대조군 (hteam)	94	8.151722	5.210929
실험군1 (aiteam)	101	7.204542	4.585274
t-검정(p-value)	0.1807		

자료: 저자작성

다음으로 대조군(hteam)과 실험군1(aiteam)의 과업2 결과물을 비교한 t-검정 결과, 실험군1(aiteam)의 과업2 결과물의 품질이 대조군보다 1% 유의수준에서 더 높았으며 과업 완료에 소요된 시간 역시 1% 유의수준에서 더 짧은 것으로 나타났다. 두 집단의 과업결과물의 품질 분포 비교는 t-검정에서의 결과와 비슷한 것으로 나타났는데, K-S 검정 결과 두 집단 분포 형태가 10% 유의수준에서 상이하며, 순위합 검정 결과 5% 유의수준에서 실험군1(aiteam)의 과업 2 결과물 품질 분포가 대조군(hteam)에 비해 더 높은 경향이 있는 것으로 나타났다.

〈표 3-15〉 과업2의 품질: 대조군(hteam)과 실험군1(aiteam)

집단	관측 수	평균(점)	표준편차
대조군 (hteam)	85	5.230588	1.19742
실험군1 (aiteam)	102	5.670588	1.008733
t-검정(p-value)	0.0080		
K-S검정(p-value)	0.080		
순위합검정(p-value)	0.015		

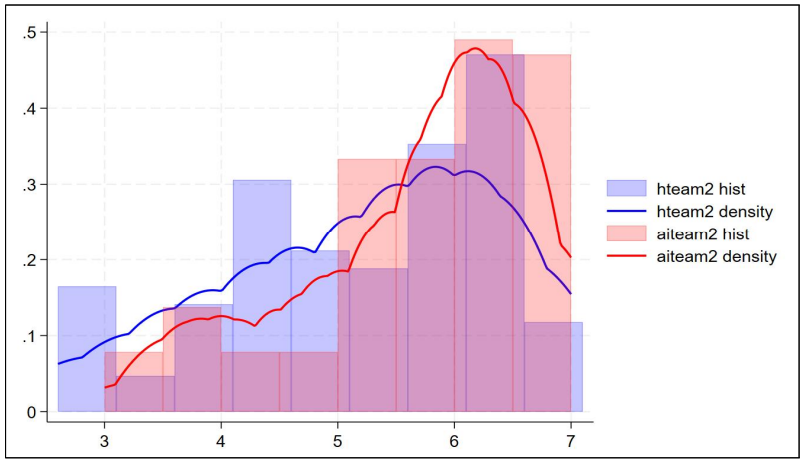
자료: 저자작성

〈표 3-16〉 과업2의 소요 시간: 대조군(hteam)과 실험군1(aiteam)

집단	관측 수	평균(분)	표준편차
대조군 (hteam)	85	16.35485	7.615256
실험군1 (aiteam)	102	12.99939	7.959433
t-검정(p-value)	0.0037		

자료: 저자작성

[그림 3-13] 과업2 결과물의 품질: 대조군(hteam) vs. 실험군1(aiteam)



자료: 저자작성

이상의 결과를 종합하면, 분업이 존재할 경우, 생성형 AI의 활용이 과업1에는 큰 영향을 미치지 않는 반면, 과업2에는 유의미하게 긍정적인 영향을 미친다는 결론에 도달하게 된다. 먼저 이 자체로 분업이 존재할 경우, 생성형 AI의 활용이 각 과업에 미치는 영향 자체가 달라진다는 결론을 도출할 수 있다. 분업이 없을 경우, 생성형 AI 활용은 과업1에만 긍정적인 영향을 주었지만, 분업이 존재할 경우 반대로 생성형 AI 활용이 과업2에만 긍정적인 영향을 주게 된다는 결론이다. 하지만 이러한 결론은 분업의 부정적인 효과를 고려하지 않는 문제가 있다. 가설 4

는 생성형 AI의 활용이 분업 효과에 영향을 미치지 않는다는 가정하에서 과업1 결과물의 유의미한 차이가 과업2 결과물 품질의 유의미한 차이로 이어진다는 것인데, 검증 결과 과업1 결과물의 품질에는 차이가 존재하지 않고, 과업2 결과물의 품질에서만 차이가 발생하였다. 이는 가설 4를 뒷받침할 수 없는 결과로, 가설 4를 도출하는 데 사용된 가정인 "생성형 AI의 활용이 분업 효과에 영향을 미치지 않는다"가 틀렸음을 시사한다. 즉, 생성형 AI의 활용은 분업의 효과에도 영향을 미칠 가능성이 있다.

먼저 과업1의 결과물 품질에서 대조군(hteam)과 실험군1(aiteam) 간의 차이가 존재하지 않는 것은 분업으로 인해 과업1 수행에서 나타나는 행태적 편향으로 설명할 수 있다. 즉, 분업으로 인한 행태적 편향, 즉 과업1의 수행이 최종 결과물이 아니며, 자신이 향후 사용할 자료가 아니라는 점에서 기인하는 불성실한 과업 수행으로 인해 실험군1(aiteam)에서의 생성형 AI 활용이 제대로 이루어지지 않아 두 집단 간 과업1 결과물의 품질에서 유의미한 차이가 나타나지 않을 수 있다. 불성실한 과업 수행이라는 행태적 편향은 두 집단 모두에서 존재하나, 이 경우 생성형 AI의 활용에 있어서 프롬프팅에 대한 고민없을 경우 생성형 AI의 생산성 향상 효과는 제한적으로 나타날 수 밖에 없다.

반면 과업2 결과물의 품질이 실험군1(aiteam)에서 더 높게 나타나는 것은 생성형 AI의 활용이 과업2 수행에 있어서 긍정적인 효과를 제공한다는 것을 시사한다. 하지만 결과 3에서 살펴보았듯이 생성형 AI 활용 그 자체로 과업2 결과물의 품질을 높이지 못한다. 실험군1(aiteam)과 대조군(hteam)의 과업2 결과물 비교와 결과 3에서의 실험군3(aiindv)과 실험군2(hindv)의 과업2 결과물 비교에서의 차이는 분업의 유무이다. 즉, 실험군1(aiteam)이 대조군(hteam)에 비하여 과업2 결과물 품질이 더 높은 이유는 생성형 AI 활용과 분업 효과 간의 상호작용으로 해석할 수 있다. 따라서 분업이 존재하는 경우, 생성형 AI의 활용이 분업의 부정적인 효과를 상쇄하는 것으로 해석할 수 있다. 두 집단의 과업1 결과물의 차이가 없는 것은 이러한 해석에 힘을 실어주는데, 과업1 결과물의 품질 차이가 없는 상태

에서 과업2 결과물의 품질은 오로지 분업의 부정적 효과와 생성형 AI 활용 여부에 의하여 결정된다. 생성형 AI의 활용이 과업2 수행에 직접적인 도움을 주지 않는다는 결과 3까지 고려하면, 대조군(hteam)과 실험군1(aiteam)의 과업2 결과물의 차이는 생성형 AI 활용이 분업의 부정적인 효과에 미치는 영향에 기인하며, 이러한 영향은 생산성을 증대하는 방향으로 작용한다. 즉, 생성형 AI 활용이 분업의 부정적인 효과를 줄여주는 방향으로 작용한다고 볼 수 있다.

결과 4 (생성형 AI가 분업에 미치는 영향1). 실험군1(aiteam)과 대조군(hteam)의 과업1 결과물의 품질은 큰 차이가 없으나, 과업2 결과물의 품질은 실험군1(aiteam)이 대조군(hteam)보다 높다. 두 집단 간에 과업1 결과물 품질이 다르지 않은 것은 분업으로 인해 발생하는 과업1에서의 행태적 편향에 기인하며, 두 집단 간 과업2 결과물 품질의 차이는 생성형 AI가 분업의 부정적인 효과를 상쇄하는 데에 기인한다.

마지막으로 가설 5의 검증은 결과 4를 더욱 뒷받침한다. 가설 5는 생성형 AI가 분업 효과에 영향을 미치지 않는다는 가정하에서 실험군1(aiteam)과 실험군3(aiindv)의 과업 결과물의 품질을 예상하였다. 실험군1(aiteam)과 실험군3(aiindv)의 비교는 분업의 효과가 존재하는 상황에서 생성형 AI 활용이 각 과업에 미치는 영향을 파악함으로써 생성형 AI 활용이 분업 효과에 미치는 영향을 조금 더 직접적으로 살펴볼 수 있다.

먼저 실험군1(aiteam)과 실험군3(aiindv)의 과업1 결과물을 비교할 때, 만약 앞서 살펴본 과업1 수행에 있어서의 행태적 편향이 존재한다면, 실험군1(aiteam)의 과업1 결과물의 품질이 실험군3(aiindv)보다 낮을 것이다. 이러한 상황에서 결과 4에서 살펴본 생성형 AI가 분업의 비효율성을 상쇄하는 효과가 나타난다면, 실험군(aiteam)의 과업2 결과물의 품질이 분업의 부정적인 효과를 상쇄하며 증가하게 되고, 결국 실험군3(aiindv)의 과업2 결과물의 품질과 큰 차이가 없게 될 것이다. 먼저 실험군1(aiteam)과 실험군3(aiindv)의 과업1 결과물을 비교한 t-검정 결과, 과업1 수행에서의 행태적 편향으로 인해 실험군1(aiteam)의 과업1 결과물

이 5% 유의수준에서 실험군3(aiindv)보다 낮게 나타난다. 단, 과업1 완료에 소요되는 시간에서는 두 집단 간 차이가 존재하지 않는다. 두 집단 간 과업결과물 품질 분포 비교에서는 K-S검정과 순위합 검정 모두에서 유의미한 차이를 확인하지 못하였다.

〈표 3-17〉 과업1의 품질: 실험군1(aiteam1)과 실험군3(aiindv)

집단	관측 수	평균(점)	표준편차
실험군1 (aiteam)	101	5.518152	1.131837
실험군3 (aiindv)	87	5.842912	.5129816
t-검정(p-value)	0.0106		
K-S검정(p-value)	0.706		
순위합검정(p-value)	0.2531		

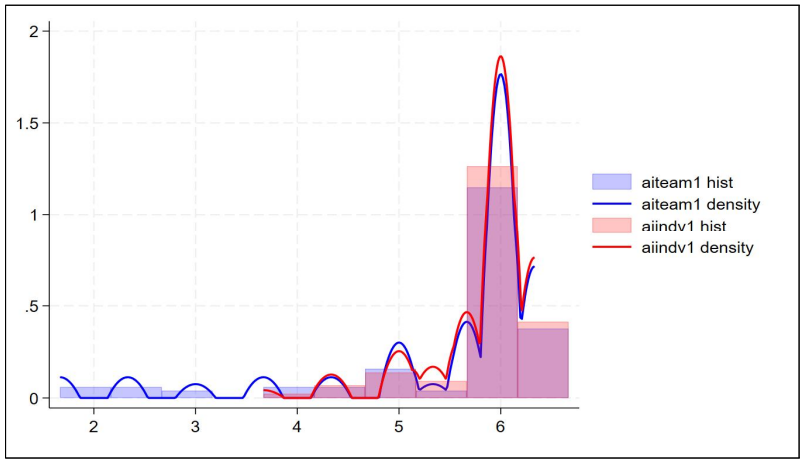
자료: 저자작성

〈표 3-18〉 과업1의 소요 시간: 실험군1(aiteam1)과 실험군3(aiindv)

집단	관측 수	평균(분)	표준편차
실험군1 (aiteam)	101	7.204542	4.585274
실험군3 (aiindv)	87	7.369988	4.408733
t-검정(p-value)	0.8014		

자료: 저자작성

[그림 3-14] 과업1 결과물의 품질: 실험군1(aiteam) vs. 실험군3(aiindv)



자료: 저자작성

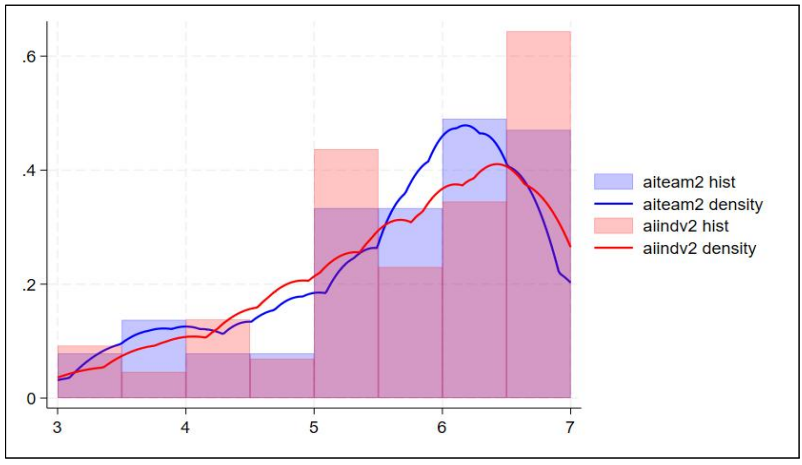
다음으로 실험군1(aiteam)과 실험군3(aiinvd)의 과업2 결과물을 비교한 t-검정 결과, 두 집단 간의 과업2 결과물의 품질 및 완료에 소요된 시간의 차이는 통계적으로 유의미하지 않은 것으로 나타났다. 과업결과물 품질 분포에서도 K-S 검정과 순위합 검정 결과 유의미한 차이를 확인하지 못하였다.

〈표 3-19〉 과업2의 품질: 실험군1(aiteam1)과 실험군3(aiindv)

집단	관측 수	평균(점)	표준편차
실험군1 (aiteam)	102	5.670588	1.008733
실험군3 (aiindv)	87	5.700529	1.024125
t-검정(p-value)	0.6618		
K-S검정(p-value)	0.873		
순위합검정(p-value)	0.5755		

자료: 저자작성

[그림 3-15] 과업2 결과물의 품질: 실험군1(aiteam) vs. 실험군3(aiindv)



자료: 저자작성

〈표 3-20〉 과업2의 소요 시간: 실험군1(aiteam1)과 실험군3(aiindv)

집단	관측수	평균(분)	표준 편차
실험군1 (aiteam)	102	12.99939	7.959433
실험군3 (aiindv)	87	13.64013	7.846716
t-검정(p-value)	0.5790		

자료: 저자작성

결과 4에서 살펴보았듯 분업이 존재하는 경우, 생성형 AI의 활용은 분업의 부정적인 효과를 상쇄할 수 있다. 실험군1(aiteam)과 실험군3(aiindv)의 비교는 이러한 해석에 힘을 실어주는데, 분업의 부정적인 효과 중 하나인 행태적 편향으로 인해 실험군1(aiteam)의 과업 1 결과물의 품질이 실험군3(aiindv)에 비하여 더 낮은 반면, 생성형 AI 활용으로 인해 행태적 편향을 포함한 분업의 부정적 영향이 상쇄되어 종국적으로는 실험군1(aiteam)과 실험군3(aiindv)의 과업 2 결과물에는 유의미한 차이가 나타나지 않는 것을 확인할 수 있다.

4. 소결

본 절에서는 실험을 통해 수집한 데이터를 분석하였다. 데이터 분석 결과 첫째, 분업의 부정적인 효과를 확인하였다. 기존 연구들에서 언급되었던 정보손실 또는 정보해석 비용 등 분업의 부정적인 효과에 더하여, 과업1의 수행에서 발생하는 행태적 편향으로 인한 추가적인 부정적 효과를 확인하였다. 즉, 분업으로 인해서 자연적으로 발생하는 정보의 손실과 전달된 정보를 해석하면서 발생하는 오류 등으로 인한 비용에 더하여, 과업1이 과업의 최종 과업 결과물이 아니며 자신이 추후에 사용할 결과물이 아니기 때문에 과업1의 수행에서 더 낮은 수준의 노력만을 기울이게 되는 행태적 편향으로 인한 부정적인 효과를 추가적으로 확인하였다. 이와 같은 행태적 편향은 분업이 존재하는 모든 실험 상황에서 일관적으로 관측되었다. 즉, 대조군(hteam)과 실험군2(hindv), 대조군(hteam)과 실험군1(aiteam), 실험군1(aiteam)과 실험군3(aiindv)의 과업1 결과물의 비교에서 관측되었다.

〈표 3-21〉 결과 요약

결과	비교 집단 1	비교 집단 2
분업의 부정적 효과		
(결과 1) 정보손실 또는 정보해석 비용	대조군(hteam) vs. 실험군2(hindv)	
	분업 ○ AI ×	분업 × AI ×
(결과 1) 과업1에서의 행태적 편향	대조군(hteam) vs. 실험군2(hindv)	
	분업 ○ AI ×	분업 × AI ×
	대조군(hteam) vs. 실험군1(aiteam)	
	분업 ○ AI ×	분업 ○ AI ○
	실험군1(aiteam) vs. 실험군3(aiindv)	
	분업 ○ AI ○	분업 × AI ○

결과	비교 집단 1	비교 집단 2
생성형 AI의 생산성 향상 효과		
(결과 2) 생성형 AI 활용에 따른 분업의 소멸로 인한 생산성 향상	대조군(hteam) vs. 실험군3(aiindv)	
	분업 ○ AI ×	분업 × AI ○
(결과 3) 생성형 AI와 각 과업의 생산성 향상 효과	실험군2(hindv) vs 실험군3(aiindv)	
	분업 × AI ×	분업 × AI ○
(결과 4) 생성형 AI 활용으로 인한 분업의 부정적 효과 상쇄	대조군(hteam) vs. 실험군1(aiteam)	
	분업 ○ AI ×	분업 ○ AI ○
	실험군1(aiteam) vs. 실험군3(aiindv)	
	분업 ○ AI ○	분업 × AI ○

자료: 저자작성

둘째, 생성형 AI의 활용으로 인해 분업이 소멸되고, 이로 인한 생산성 증가 효과를 확인하였다. 분업이 존재하는 대조군(hteam)의 상황과 생성형 AI가 도입되어 모든 과업에 활용되며 분업이 사라지는 상황인 실험군3(aiindv)을 비교한 결과, 모든 과업 결과물의 품질이 향상되고 과업에 소요되는 시간 역시 단축되는 것을 확인하였다. 먼저 분업의 소멸은 분업으로 인해 발생하였던 과업1 수행에서의 행태적 편향의 제거와, 과업1 수행에서 생성형 AI 활용으로 인한 생산성 향상이 복합적으로 나타나 과업1 결과물의 품질을 향상시킨다. 또한 분업의 소멸로 인해 분업의 부정적인 효과인 정보손실 또는 정보해석 비용이 사라지게 되어 과업2 결과물의 품질 역시 상승하는 것을 확인할 수 있었다. 두 과업 모두 생성형 AI의 활용에 따라 과업 완료에 소요되는 시간이 단축되는 것 역시 확인하였다.

셋째, 생성형 AI가 생산성을 높일 수 있는 과업과 그렇지 않은 과업의 차이를 다시 한번 확인하였다. 기존 문헌에 따르면, 본 실험에서 과업1(정보의 요약)은 비교적 단순하고 복잡하지 않아 생성형 AI가 생산성을 쉽게 높일 수 있는 반면, 과업2(구매설득 이메일 작성)는 창의성과 정보의 재조합이 필요한 복잡한 과업으로

생성형 AI로 인한 생산성 향상 효과가 제한적이다. 본 실험에서는 생성형 AI 활용이 각 과업에 미치는 영향을 분업이 존재하지 않는 환경에서 살펴보기 위하여 실험군2(hindv)와 실험군3(aiindv)의 과업 결과물을 비교하였는데, 비교 결과 기존 문헌에서처럼 과업1의 경우 결과물의 품질 향상 및 과업 완료에 소요되는 시간의 단축을 통해 생성형 AI의 생산성 향상 효과를 확인하였지만, 과업2의 경우 생성형 AI의 활용이 과업 결과물의 품질 향상 또는 소요 시간 단축으로 이어지지 않아 생산성 향상 효과가 제한적임을 재확인하였다.

마지막으로 생성형 AI의 활용이 분업의 부정적 효과를 상쇄하는 것을 확인하였다. 생성형 AI의 활용이 곧바로 분업의 소멸로 이어지지 않을지라도, 분업을 유지한 상태에서 생성형 AI를 활용할 경우 정보손실 또는 정보해석 비용에 따른 부정적인 효과가 일정 부분 상쇄됨을 대조군(hteam)과 실험군1(aiteam)의 비교를 통해 확인하였다. 하지만, 생성형 AI의 활용이 분업으로 인해 발생하는 과업1 수행에서의 행태적 편향을 낮추지는 못하는 것을 확인하였다. 이러한 결과는 실험군1(aiteam)과 실험군3(aiindv)의 과업 결과물 비교를 통해서 재확인되는데, 분업으로 인한 과업1 수행에서의 행태적 편향으로 인해 실험군1(aiteam)의 과업1 결과물의 품질이 실험군3(aiindv)의 과업1 결과물의 품질보다 낮지만, 분업으로 인한 부정적인 효과와 과업1의 낮은 품질에도 불구하고 생성형 AI 활용은 실험군1(aiteam)의 과업2 결과물의 품질을 높여 분업으로 인한 과업1 수행에서의 행태적 편향 및 정보손실 등의 부정적인 효과가 없는 상태인 실험군3(aiindv)의 과업2 결과물의 품질과 큰 차이가 존재하지 않는 것을 확인하였다. 이러한 결과는 생성형 AI의 활용이 과업2 수행에서 생산성 향상에는 직접적으로 도움이 되지 않는다는 결과를 고려할 때, 생성형 AI가 분업의 부정적인 효과를 상쇄한다고 해석할 수 있다.

제 4 장 결 론

생성형 AI의 활용이 곧 생산성 증가로 이어진다는 결과는 다수의 연구에서 이미 밝혀져 있다. 하지만 이러한 생산성 향상이 어떠한 경로를 통해 이루어지는지에 대해서는 아직 구체적인 연구가 없는 실정이다. 본 연구는 생성형 AI의 생산성 향상 경로 중 분업과 관련된 경로에 주목하여, 실험을 설계하고 생성형 AI 활용과 분업의 상호작용을 분석하고자 하였다. 실험은 생성형 AI의 활용 여부와 분업의 유무가 상호작용할 수 있는 환경을 구현하고 분석하기 위하여 2×2 구조를 활용하였다. 즉, 생성형 AI 활용 여부와 분업의 유무에 따라 총 네 종류의 집단이 두 개로 구성된 과업을 수행하였다. 실험에 활용된 과업은 마케팅 과업으로, 과업1은 1,200자 설명문을 읽고 요약하는 과업, 과업2는 요약된 정보를 바탕으로 구매설득 이메일을 발송할 대상을 특정하고 구매설득 이메일을 작성하는 것으로, 과업1과 과업2는 분업을 통해 수행될 수도, 분업 없이 한 명의 개인이 연속적으로 수행할 수 있다. 해당 과업에서 사용된 제품은 가상의 제품을 상정하여 실험을 진행하였다. 실험은 만 19세 이상 만 60세 미만의 성인을 대상으로 온라인에서 진행되었고, 총 602명이 실험에 참여하였고, 이 중 불성실한 응답 또는 실험 내용과 관련 없는 응답을 제외한 533명의 데이터를 분석하였다. 553명의 과업1과 과업2 결과물은 ChatGPT를 활용하여 동일한 기준으로 품질을 측정하였고, 과업에 소요된 시간은 온라인 실험 시 각 과업을 수행하는 페이지에 접근한 시간과 해당 페이지에서 벗어난 시간의 차이를 구하여 측정하였다.

총 553명의 데이터를 분석한 결과, 본 연구에서는 ① 생성형 AI의 활용에 따른 분업의 소멸에 의한 생산성 향상과 ② 분업이 소멸하지 않더라도, 생성형 AI의 활용이 분업의 부정적인 효과를 상쇄하여 생산성을 향상시키는 두 가지 생산성 향상 경로를 확인하였다. 첫 번째 생산성 향상 경로는 생성형 AI의 활용에 따라 종전에는 분업을 통해서만 수행할 수 있었던 과업이 분업 없이 한 명이 모두 수행할

수 있게 될 경우, 분업이 생산성에 미치는 부정적인 효과가 제거되고 이로 인해 생산성이 향상될 수 있다. 두 번째 생산성 향상 경로는 분업이 그대로 유지되지만 각 세부과업에서 생성형 AI가 활용될 경우, 분업이 생산성에 미치는 부정적인 효과를 상쇄시킴으로써 생산성을 향상시킨다. 흥미롭게도 두 생산성 향상 경로를 통해 증가된 생산성 향상의 증분은 크게 차이가 나지 않는다. 이는 노동시장에 함의를 가지는데, 단순히 AI가 노동을 대체하여 증가하는 생산성 증분과 AI가 노동을 대체하지 않고 보완하여 증가하는 생산성 증분이 크게 다르지 않을 수 있음을 의미한다. 하지만 AI 활용 비용과 임금에 대한 논의가 포함되어 있지 않아 해석에 주의가 필요하다.

참 고 문 헌

[국내 문헌]

민대홍(2024. 6), “생성형 AI의 생산선성분석 실험연구”, KISDI Perspectives, 6월호, 정보통신정책연구원.

《대한민국 정책브리핑》(2024. 4. 4), “‘AI 일상화’에 올해 7100억 원 투입…AI로 건강관리·돌봄 등 지원”, https://www.korea.kr/news/policyNewsView.do?newsId=148927861.

[해외문헌]

Brynjolfsson, E., Li, D., & Raymond, L. (2025). “Generative AI at work.” *The Quarterly Journal of Economics*, 140(2), 889-942. Brynjolfsson, E., Li, D., & Raymond, L. (2025). “Generative AI at work.” *The Quarterly Journal of Economics*, 140(2), 889-942

Choi, J. H. and Schwarcz, D.(2024). “Ai assistance in legal analysis: An empirical study.” *J. Legal Educ.*, 73, 384.

Dessein, W. and Santos, T.(2006). “Adaptive organizations.” *Journal of Political Economy*, 114(5), 956-995.

Dessein, W., Galeotti, A. and Santos, T.(2016). “Rational inattention and organizational focus.” *American Economic Review*, 106(5), 152-156.

Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., ... and Lakhani, K. R.(2023). “Navigating the jagged technological frontier: Field experimental evidence of

the effects of AI on knowledge worker productivity and quality." Harvard Business School Technology & Operations Mgt. Unit Working Paper, (24-013).

Doshi, A. R. and Hauser, O. P.(2024). "Generative AI enhances individual creativity but reduces the collective diversity of novel content." *Science Advances*, 10(28), eadn5290.

Garicano, L.(2000). "Hierarchies and the organization of knowledge in production." *Journal of Political Economy*, 108(5), 874-904.

Haslberger, M., Gingrich, J. and Bhatia, J.(2023). "No great equalizer: experimental evidence on AI in the UK labor market." Available at SSRN: https://ssrn.com/abstract=4594466

Jia, N., Luo, X., Fang, Z. and Liao, C.(2024). "When and how artificial intelligence augments employee creativity." *Academy of Management Journal*, 67(1), 5-32.

Kim, D. G. and Moon, A.(2025). "From helping hand to stumbling block: the ChatGPT paradox in competency experiment." *Applied Economics Letters*, 32(18), 2669-2673.

Peng, S., Kalliamvakou, E., Cihon, P. and Demirer, M.(2023). "The impact of ai on developer productivity: Evidence from github copilot." arXiv preprint arXiv:2302.06590.

Noy, S. and Zhang, W.(2023). "Experimental evidence on the productivity effects of generative artificial intelligence." *Science*, 381(6654), 187-192.

Nie, A., Chandak, Y., Suzara, M., Malik, A., Woodrow, J., Peng, M., ... and Piech, C.(2024). "The GPT surprise: offering large language model chat in a massive coding class reduced engagement but

increased adopters exam performances." arXiv preprint arXiv:2407.09975.

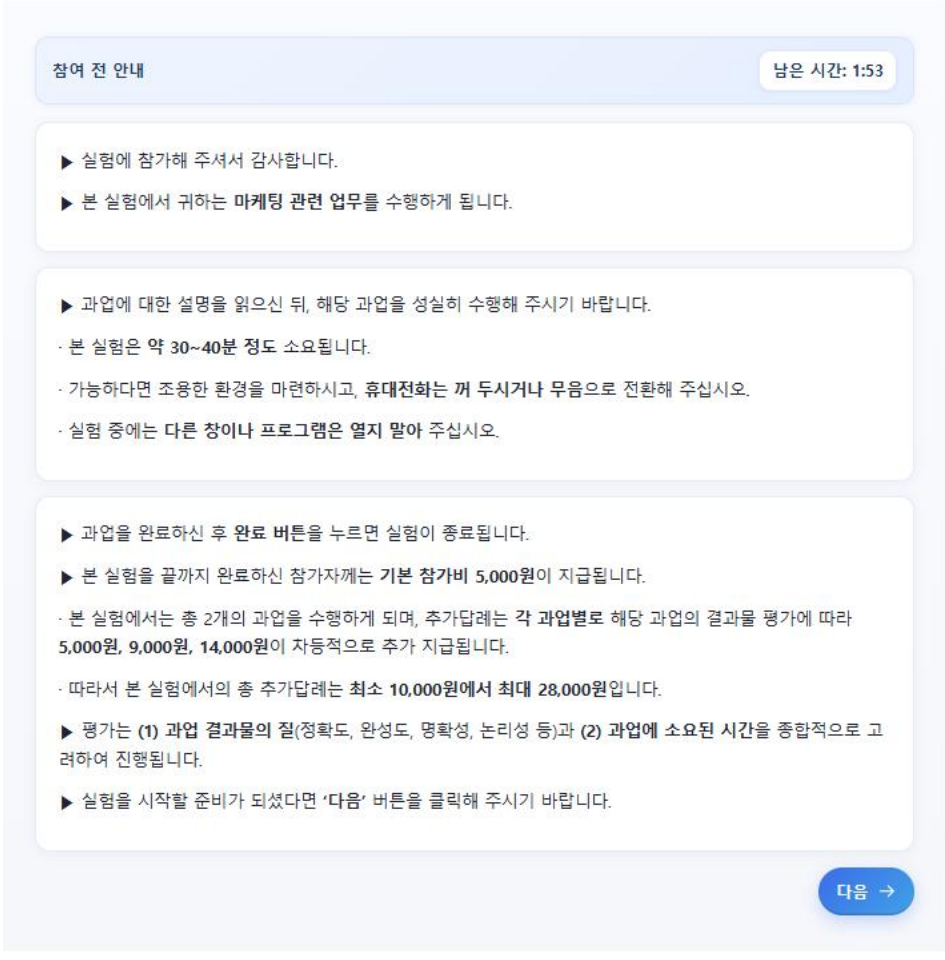
Radner, R.(1993). "The organization of decentralized information processing." *Econometrica*, 61(5), 1109-1146.

Roldán-Monés, T.(2024). "When GenAI increases inequality: evidence from a university debating competition." Working Paper 24/09, EsadeEcPol and London School of Economics, Barcelona. Working Paper, version of 26 August.

Smith, A.(1776). *An Inquiry into the Nature and Causes of the Wealth of Nations.* London: W. Strahan and T. Cadell.

부 록

[실험군2(hindv) 실험화면]



업무 안내 남은 시간: 4:58

▶ 귀하께서는 새로 출시되는 **스마트 족욕기 오르비스파(OrbiSpa)**의 마케팅 관련 업무를 수행하게 됩니다.

▶ 귀하께서 수행할 업무는 해당 신제품의 **구매 설득 이메일**을 작성하는 것입니다.

- 구매설득 이메일 작성은 2개의 세부과업으로 이루어져있습니다.

과업(1) 제품 설명문을 바탕으로, 구매 설득 이메일 작성을 위한 기초자료 요약 및 작성

과업(2) 구매층(타겟층) 특정 및 구매 설득 이메일 작성

다음 →

참여 전 안내 남은 시간: 4:58

▶ 다음 페이지에 제시되는 1,200자 분량의 설명문을 읽고, 지시사항에 따라 자료를 요약해 주시기 바랍니다.

· 본 과업을 수행하는 동안, **다른 프로그램이나 창 사용은 엄격히 제한**됩니다. 추후 검증을 통해 다른 프로그램 사용 흔적이 발견될 경우, **추가 보상은 지급되지 않습니다.**

▶ 아래에 요구된 항목을 작성 규칙에 따라 작성하십시오. **(다음 버튼을 누르면 작성페이지로 넘어갑니다.)**

· 아래 요구 항목과 작성규칙은 다음 작성페이지에서도 보실 수 있습니다.

▶ **요구 항목**

- 제품의 기능(기기가 수행하는 동작 또는 기술적 특징) **3가지**
- 구매유도 요인(소비자에게 구매를 유도할 수 있는 제품의 장점) **3가지**
- 구매저해 요인(소비자가 구매를 망설이게 되는 이유) **3가지**

▶ **작성 규칙**

- 각 항목은 **한 문장**으로 간단히 요약할 것
- 문서에 없는 내용을 임의로 작성하지 말 것
- 문장을 그대로 복사하지 말고, **자신의 표현으로 요약**할 것

다음 →

과업 1 · 설명 요약
남은 시간: 19:59
설명문
오르비스파(OrbiSpa)"는 집에서도 스파의 편안함을 경험할 수 있도록 설계된 개인용 스마트 족욕기다. 제조사는 이 제품을 단순한 온수 기기가 아닌, 사용자의 일상 속 피로 회복과 심리적 안정을 돕는 '홈 스파 플랫폼'으로 정의한다. 빠르게 성장하는 웰니스 시장 속에서 오르비스파는 하루의 긴장을 풀고 스스로를 돌보는 시간을 제안하며, 기존 족욕기의 단순한 온열 기능을 넘어선 새로운 생활가전으로 주목받고 있다.
오르비스파는 최신 기술을 도입하여 제품의 가치를 극대화하였다. 내장된 센서가 사용자의 습관과 선호를 학습해 자동으로 온도와 시간을 조절하며, 별도의 조작 없이도 일정한 족욕 환경을 제공한다. 스마트폰 앱과 연동되어 예약 가동
▶ 요구 항목
제품의 기능(기기가 수행하는 동작 또는 기술적 특징) 3가지
구매유도 요인(소비자에게 구매를 유도할 수 있는 제품의 장점) 3가지
구매저해 요인(소비자가 구매를 망설이게 되는 이유) 3가지
▶ 작성 규칙
각 항목은 한 문장으로 간단히 요약할 것
문서에 없는 내용을 임의로 작성하지 말 것
문장을 그대로 복사하지 말고, 자신의 표현으로 요약할 것
메모장
(아이디어를 메모하거나 초안을 작성해두세요. 메모는 제출되지 않습니다.)

구분	입력란
기능1	
기능2	
기능3	
구매유도 요인1	
구매유도 요인2	
구매유도 요인3	
구매저해 요인1	
구매저해 요인2	
구매저해 요인3	

다음 →

다음 과업 안내 남은 시간: 4:59

- 다음 과업은 **구매층(타겟층) 특정**입니다.
- 본 과업을 수행하는 동안 전 화면에서 작성하신 **요약자료가 제공됩니다.**
- 다음 버튼을 누르시면, 다음 과업으로 넘어갑니다.
- 다음 버튼을 누르신 후에는 이전 화면으로 돌아갈 수 없습니다.

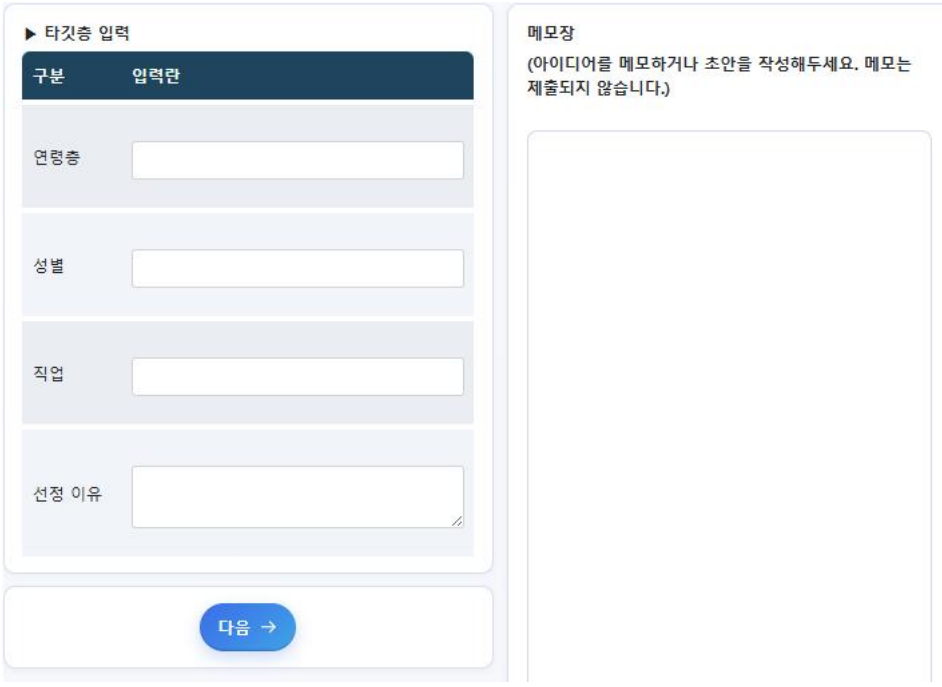
다음 →

과업 2-1 · 구매층(타깃층) 특정 남은 시간: 19:59

- 다음은 귀하가 설명문을 바탕으로 요약한 **스마트 족욕기 오르비스파(OrbiSpa)**의 기능 3가지, 구매유도요인 3가지, 그리고 구매저해요인 3가지 입니다.
- 귀하가 작성하신 자료를 바탕으로, 아래 지시사항에 따라 (1) 구매설득 이메일을 발송할 **구매층을 특정**하고, (2) 해당 구매층을 대상으로 제품을 소개하고 구매를 설득하는 메시지를 작성하십시오.
- 본 과업을 수행하는 동안, **다른 프로그램이나 창을 사용하는 행위는 엄격히 제한**됩니다. 추후 검증을 통해 타 프로그램 사용 흔적이 확인될 경우, **추가 보상은 지급되지 않습니다.**

과업 1의 결과물

구분	입력값
기능1	—
기능2	—
기능3	—
구매유도요인1	—
구매유도요인2	—
구매유도요인3	—
구매저해요인1	—
구매저해요인2	—
구매저해요인3	—


▶ 타깃층 입력
구분
입력란
연령층
성별
직업
선정 이유
다음 →
메모장
(아이디어를 메모하거나 초안을 작성해두세요. 메모는 제출되지 않습니다.)

과업 2-2 · 이메일 작성 남은 시간: 19:59

▶ 아래 자료를 바탕으로, 특정한 구매층을 대상으로 제품을 소개·설득하는 이메일을 작성하세요.

▶ 본 과업을 수행하는 동안, **다른 프로그램이나 창을 사용하는 행위는 엄격히 제한**됩니다. 추후 검증을 통해 타 프로그램 사용 흔적이 확인될 경우, **추가 보상은 지급되지 않습니다.**

과업 1의 결과물

구분	입력란
기능1	—
기능2	—
기능3	—
구매유도요인1	—
구매유도요인2	—
구매유도요인3	—
구매저해요인1	—
구매저해요인2	—
구매저해요인3	—

과업 2-1 · 구매층(세그먼트) 결과물

——— — —

구분	입력란
연령층	—
성별	—
직업	—
선정 이유	—

▶ 작성 요구 사항

- 이메일 제목(Subject line): **60자 이내**
- 이메일 본문(Body): **120~150 단어**
- 주요 장점·고객 가치가 드러나야 함
- **구매저해 요인을 언급**하고 이를 상쇄할 수 있는 이유/근거/논리 제시
- 주어진 정보만 사용(새 정보 창작 금지)

이메일 제목

제목 길이: 0/60자

이메일 본문

현재 단어 수: 0 (권장 120–150)

제출 →

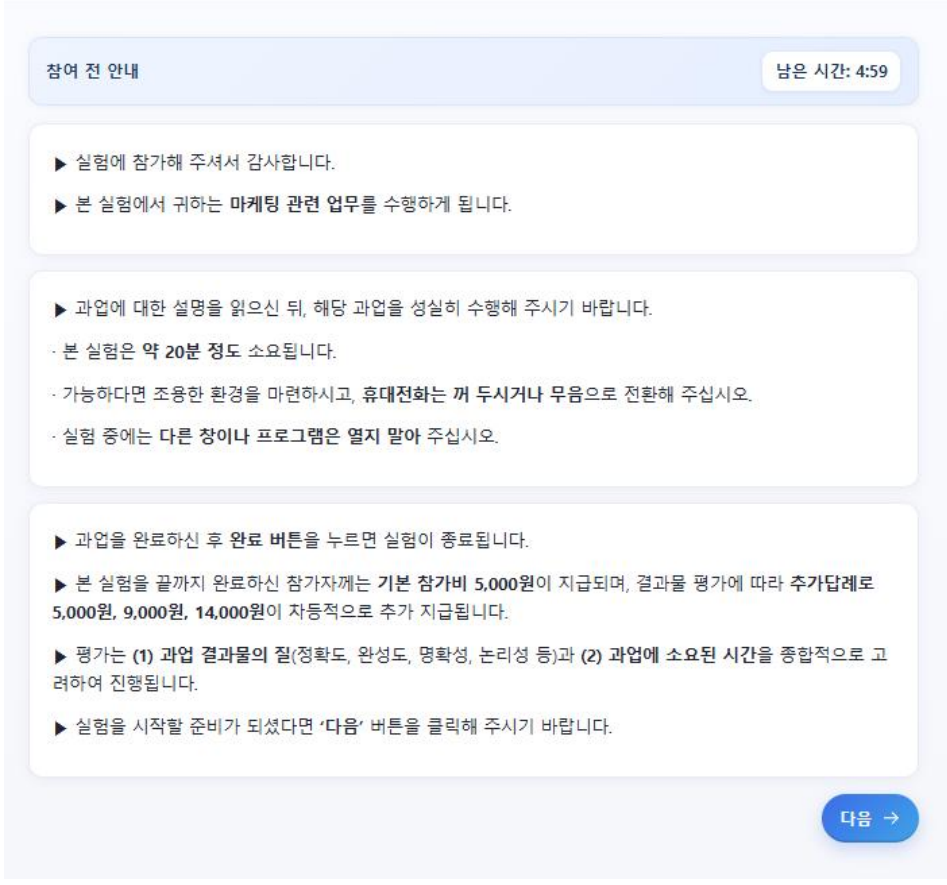
메모장

(아이디어를 메모하거나 초안을 작성해두세요. 메모는 제출되지 않습니다.)

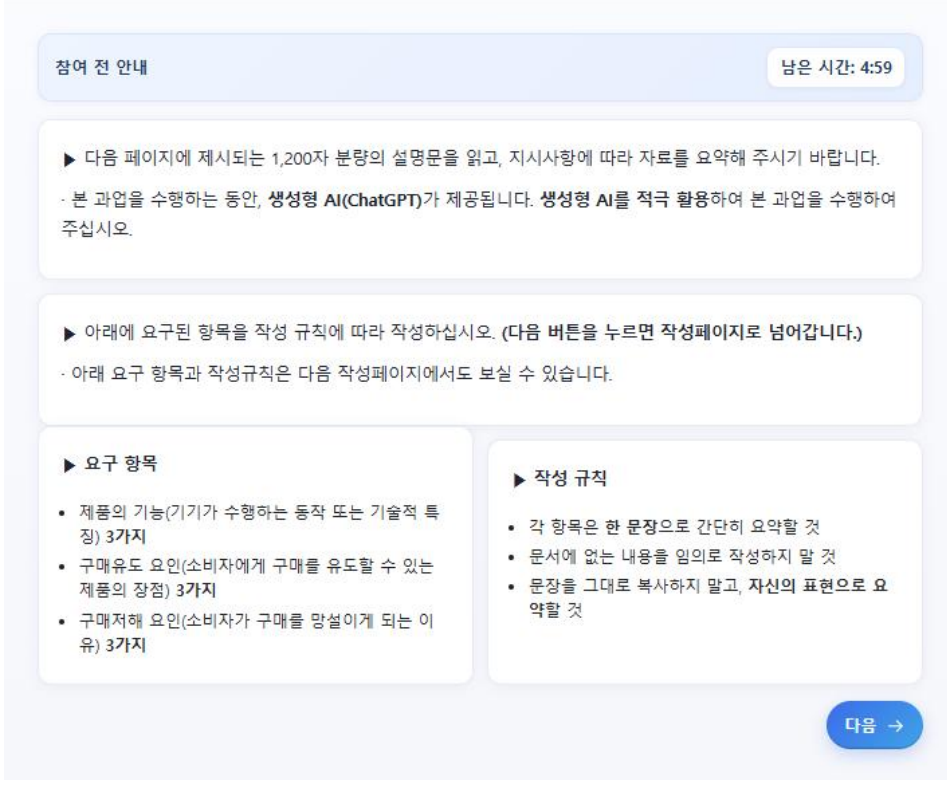
[실험군3(aiindv) 실험화면]

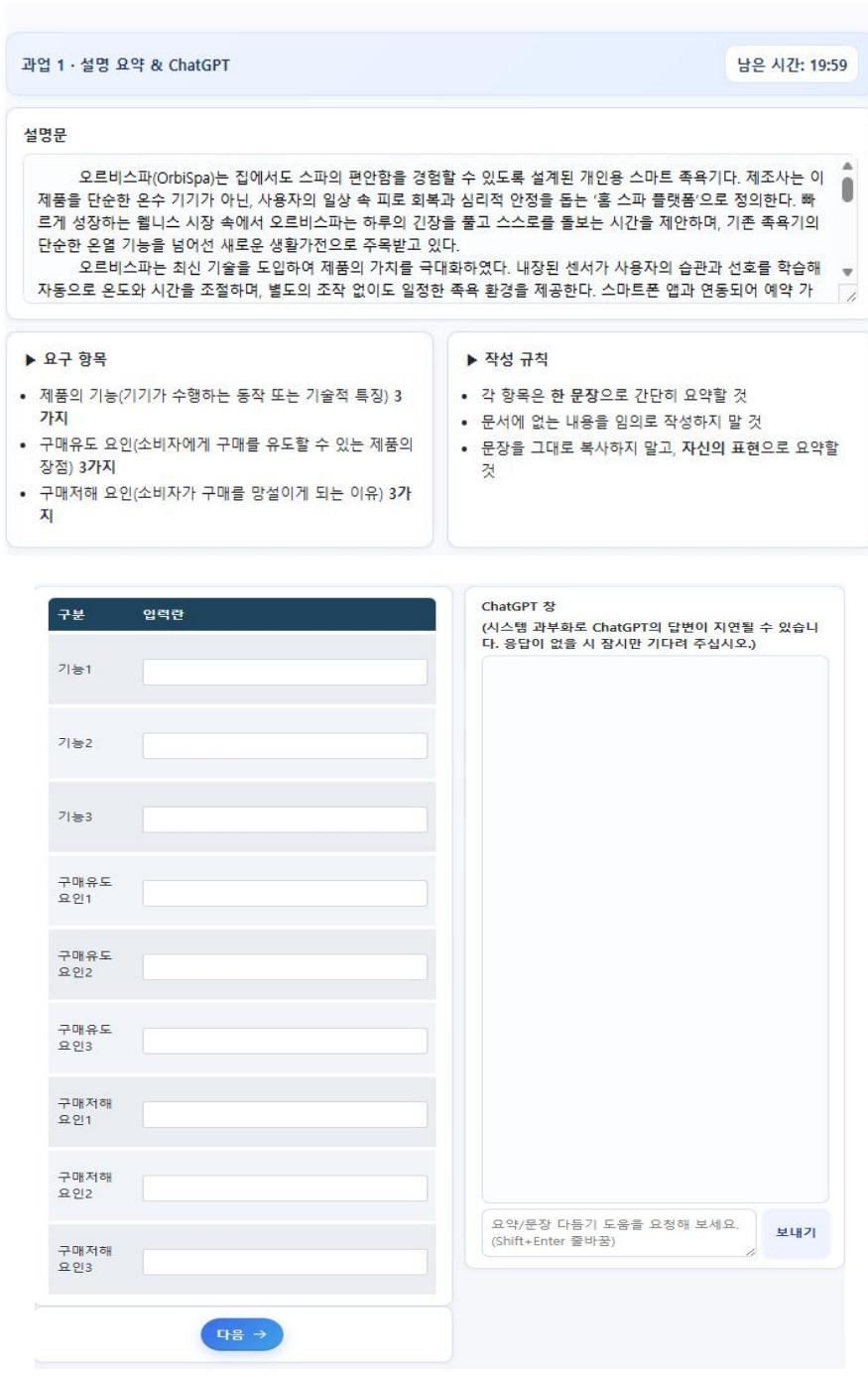
실험군3(aiindv)의 실험화면은 실험군2(hindv)와 전체적으로 동일하나, 과업 수행 시 메모장 대신 ChatGPT를 사용할 수 있는 채팅창이 제공되었으며, 과업 수행시 “본 과업을 수행하는 동안, 다른 프로그램이나 창을 사용하는 행위는 엄격히 제한됩니다. 추후 검증을 통해 타 프로그램 사용 흔적이 확인될 경우, 추가 보상은 지급되지 않습니다.” 대신 “본 과업을 수행하는 동안, 생성형 AI(ChatGPT)가 제공됩니다. 생성형 AI를 적극 활용하여 본 과업을 수행하여 주십시오.” 라는 문구가 사용됨.

[실험군1(aiteam) 실험화면 과업1 수행 실험]



업무 안내
남은 시간: 4:59
▶ 귀하께서는 새로 출시되는 스마트 족욕기 오르비스파(OrbiSpa)의 마케팅 관련 업무를 수행하게 됩니다.
▶ 귀하께서 수행할 업무는 새로운 제품에 대한 설명문을 읽고, 추후 광고 이메일 작성을 위한 기초자료를 작성하는 것입니다.
▶ 귀하께서 작성한 기초자료는 추후 광고 이메일 작성을 담당하는 다른 실험 참가자에게 전달되어 활용됩니다. 이 점을 고려하시고 과업을 수행해 주시기 바랍니다.
다음 →


참여 전 안내
남은 시간: 4:59
▶ 다음 페이지에 제시되는 1,200자 분량의 설명문을 읽고, 지시사항에 따라 자료를 요약해 주시기 바랍니다.
· 본 과업을 수행하는 동안, 생성형 AI(ChatGPT)가 제공됩니다. 생성형 AI를 적극 활용하여 본 과업을 수행하여 주십시오.
▶ 아래에 요구된 항목을 작성 규칙에 따라 작성하십시오. (다음 버튼을 누르면 작성페이지로 넘어갑니다.)
· 아래 요구 항목과 작성규칙은 다음 작성페이지에서도 보실 수 있습니다.
▶ 요구 항목
• 제품의 기능(기기가 수행하는 동작 또는 기술적 특징) 3가지
• 구매유도 요인(소비자에게 구매를 유도할 수 있는 제품의 장점) 3가지
• 구매저해 요인(소비자가 구매를 망설이게 되는 이유) 3가지
▶ 작성 규칙
• 각 항목은 한 문장으로 간단히 요약할 것
• 문서에 없는 내용을 임의로 작성하지 말 것
• 문장을 그대로 복사하지 말고, 자신의 표현으로 요약할 것
다음 →


과업 1 · 설명 요약 & ChatGPT
남은 시간: 19:59
설명문
오르비스파(OrbiSpa)는 집에서도 스파의 편안함을 경험할 수 있도록 설계된 개인용 스마트 족욕기다. 제조사는 이 제품을 단순한 온수 기기가 아닌, 사용자의 일상 속 피로 회복과 심리적 안정을 돕는 '홈 스파 플랫폼'으로 정의한다. 빠르게 성장하는 웰니스 시장 속에서 오르비스파는 하루의 긴장을 풀고 스스로를 돌보는 시간을 제안하며, 기존 족욕기의 단순한 온열 기능을 넘어선 새로운 생활가전으로 주목받고 있다.
오르비스파는 최신 기술을 도입하여 제품의 가치를 극대화하였다. 내장된 센서가 사용자의 습관과 선호를 학습해 자동으로 온도와 시간을 조절하며, 별도의 조작 없이도 일정한 족욕 환경을 제공한다. 스마트폰 앱과 연동되어 예약 가
▶ 요구 항목
제품의 기능(기기가 수행하는 동작 또는 기술적 특징) 3가지
구매유도 요인(소비자에게 구매를 유도할 수 있는 제품의 장점) 3가지
구매저해 요인(소비자가 구매를 망설이게 되는 이유) 3가지
▶ 작성 규칙
각 항목은 한 문장으로 간단히 요약할 것
문서에 없는 내용을 임의로 작성하지 말 것
문장을 그대로 복사하지 말고, 자신의 표현으로 요약할 것
구분
입력란
기능1
기능2
기능3
구매유도 요인1
구매유도 요인2
구매유도 요인3
구매저해 요인1
구매저해 요인2
구매저해 요인3
다음 →
ChatGPT 창
(시스템 과부화로 ChatGPT의 답변이 지연될 수 있습니다. 응답이 없을 시 잠시만 기다려 주십시오.)
요약/문장 다듬기 도움을 요청해 보세요. (Shift+Enter 줄바꿈)
보내기

[실험군1(aiteam) 실험화면 과업2 수행 실험]

과업 2-1 · 구매층(타깃층) 특정 남은 시간: 19:58

▶ 아래에 제공되는 자료는 새로 출시되는 스마트 족욕기 오르비스파(OrbiSpa)의 **기능**(기기가 수행하는 동작 또는 기술적 특징) 3가지, **구매유도 요인**(소비자에게 구매를 유도할 수 있는 제품의 장점) 3가지, 그리고 **구매저해 요인**(소비자가 구매를 망설이게 되는 이유) 3가지입니다

▶ 아래 자료를 바탕으로, 아래 지시사항에 따라 (1) 구매설득 이메일을 발송할 **구매층을 특정**하고, (2) 해당 구매층을 대상으로 제품을 소개하고 구매를 설득하는 메시지를 작성해 주십시오.

▶ 본 과업을 수행하는 동안, 제공되는 **생성형 AI(ChatGPT)**를 적극 활용하여 본 과업을 수행해 주십시오.

- ! 아래에 제공되는 정보는 귀하와 **무작위로 매칭된 실험참가자**가 1,200자 분량의 제품 설명문을 읽고 **요약한 자료**입니다.
- ! 귀하와 매칭된 실험참가자의 **업무역량에 따라 제공되는 정보의 질이 상이**하며, 간혹 **모든 항목이 작성되지 않았을 수도** 있습니다.
- ! 귀하의 **추가 답례를 위한 과업결과물 평가**는 귀하에게 제공된 **요약자료 품질을 반영하여 공정하게 조정**되오니, **현재 제공된 정보만을 활용**하여 과업을 **성실히 수행**해주시기 바랍니다.

과업 1의 결과물 (아래 결과물은 복사 및 붙여넣기가 가능합니다.)

구분	입력란
기능1	자동 온도 및 시간 조절 : 내장된 센서가 사용자의 사용자의 취향과 필요에 따라 별도의 조작 없이도 온도와 시간을 조절하여 맞춤형 족욕 환경을 제공합니다
기능2	스마트폰 앱 연동: 스마트폰 앱과 연결되어 사용자가 원격으로 조작 및 예약이 가능하며, 사용 기록을 통해 개인의 사용 패턴도 확인 가능합니다.
기능3	UV-C LED 살균 기능: 사용 후에는 UV-C LED가 자동으로 작동하여 내부를 살균함으로써 위생을 유지할 수 있도록 돕습니다.
구매유도요인1	홈 스파 경험: 일상 속에서도 고급 스파처럼 즐길 수 있는 경험을 제공하여 스트레스 해소 및 피로 회복을 돕습니다.
구매유도요인2	편리한 관리: 자동 온도 조절 및 순환 기능으로 추가적인 물 교체가 필요 없으며, 장시간 일정한 온도가 유지되어 관리가 간편합니다.
구매유도요인3	내구성과 청결 유지: 재활용 합성수지 소재로 제작되어 내구성이 높고, 살균 기능이 있어 위생에 민감한 사용자에게 적합합니다.
구매저해요인1	제품 가격이 일반 족욕기보다 높은 편으로 진입 장벽이 될 수 있습니다.
구매저해요인2	수조 용량이 6리터로 협소한 공간에서는 사용과 보관에 불편함을 줄 수 있습니다.
구매저해요인3	살균 과정에서 발생하는 소음과 전력 소모, 전용 솔트 캡슐의 지속적인 구매가 소비자에게 부담이 될 수 있습니다.

▶ 위 자료를 바탕으로
구매설득 이메일을 발송할 구매층
을 특정해 주십시오.

▶ 작성 예시

- **연령층**: [예: 20대, 30대, 40대 이상 등]
- **성별**: 남성 / 여성 / 성별 무관
- **직업**: [예: 대학생, 직장인, 자영업자 등]
- **선택 이유** (1~2문장): 왜 이 세그먼트가 이 제품에 관심이 있을 것 같은지 간단히 설명

구분	입력란
연령층	
성별	
직업	
선택이유	

다음 →

ChatGPT 창

(시스템 과부화로 ChatGPT의 답변이 지연될 수 있습니다. 응답이 없을 시 잠시만 기다려 주십시오.)

예) 구매설득이메일을 보낼 구매층을 특정해줘 (Shift+Enter 줄바꿈)

보내기

과업 2-2 · 이메일 작성 남은 시간: 19:59

▶ 아래 자료를 바탕으로, 특정한 구매층을 대상으로 제품을 소개·설득하는 이메일을 작성하세요.

▶ 본 과업을 수행하는 동안, **생성형 AI(ChatGPT)**가 제공됩니다. **생성형 AI를 적극 활용하여 본 과업을 수행하여 주십시오.**

- ! 아래에 제공되는 정보는 귀하와 **무작위로 매칭된 실험참가자**가 1,200자 분량의 제품 설명문을 읽고 **요약한 자료**입니다.
- ! 귀하와 매칭된 실험참가자의 **업무역량에 따라 제공되는 정보의 질이 상이**하며, 간혹 **모든 항목이 작성되지 않았을 수도** 있습니다.
- ! 귀하의 **추가 답례를 위한 과업결과물 평가**는 귀하에게 제공된 **요약자료 품질을 반영하여 공정하게 조정**되오니, **현재 제공된 정보만을 활용**하여 과업을 **성실히 수행**해주시기 바랍니다.

과업 1의 결과물

구분	입력란
기능1	자동 온도 및 시간 조절 : 내장된 센서가 사용자의 사용자의 취향과 필요에 따라 별도의 조작 없이도 온도와 시간을 조절하여 맞춤형 족욕 환경을 제공합니다
기능2	스마트폰 앱 연동: 스마트폰 앱과 연결되어 사용자가 원격으로 조작 및 예약이 가능하며, 사용 기록을 통해 개인의 사용 패턴도 확인 가능합니다.
기능3	UV-C LED 살균 기능: 사용 후에는 UV-C LED가 자동으로 작동하여 내부를 살균함으로써 위생을 유지할 수 있도록 돕습니다.
구매유도요인1	홈 스파 경험: 일상 속에서도 고급 스파처럼 즐길 수 있는 경험을 제공하여 스트레스 해소 및 피로 회복을 돕습니다.
구매유도요인2	편리한 관리: 자동 온도 조절 및 순환 기능으로 추가적인 물 교체가 필요 없으며, 장시간 일정한 온도가 유지되어 관리가 간편합니다.
구매유도요인3	내구성과 청결 유지: 재활용 합성수지 소재로 제작되어 내구성이 높고, 살균 기능이 있어 위생에 민감한 사용자에게 적합합니다.
구매저해요인1	제품 가격이 일반 족욕기보다 높은 편으로 진입 장벽이 될 수 있습니다.
구매저해요인2	수조 용량이 6리터로 협소한 공간에서는 사용과 보관에 불편함을 줄 수 있습니다.
구매저해요인3	살균 과정에서 발생하는 소음과 전력 소모, 전용 솔트 캡슐의 지속적인 구매가 소비자에게 부담이 될 수 있습니다.

과업 2-1 · 구매층(세그먼트) 결과물

—·—·— — —

구분	입력란
연령층	—
성별	—
직업	—
선정 이유	—

▶ 작성 요구 사항
이메일 제목(Subject line): 60자 이내
이메일 본문(Body): 120~150 단어
주요 장점·고객 가치가 드러나야 함
구매저해 요인을 언급하고 이를 상쇄할 수 있는 이유/근거/논리 제시
주어진 정보만 사용(새 정보 창작 금지)
이메일 제목
제목 (권장 ≤ 60자)
제목 길이: 0/60자
이메일 본문
본문 (권장 120~150 단어)
현재 단어 수: 0 (권장 120–150)
제출 →
ChatGPT 창
(시스템 과부화로 ChatGPT의 답변이 지연될 수 있습니다. 응답이 없을 시 잠시만 기다려 주십시오.)
예) 위 세그먼트에 맞춘 제목/도입부를 제안해줘 (Shift+Enter 줄바꿈)
보내기

[대조군(hteam) 실험화면]

대조군(hteam) 실험화면은 실험군1(aiteam)과 동일하나, 과업 수행 시 Chat GPT 창이 아닌 메모장이 제공되었으며, “본 과업을 수행하는 동안, 생성형 AI(ChatGPT)가 제공됩니다. 생성형 AI를 적극 활용하여 본 과업을 수행하여 주십시오.” 대신 “본 과업을 수행하는 동안, 다른 프로그램이나 창을 사용하는 행위는 엄격히 제한됩니다. 추후 검증을 통해 타 프로그램 사용 흔적이 확인될 경우, 추가 보상은 지급되지 않습니다.”라는 문구가 사용됨.

Abstract

Daehong Min

■ Title

Productivity gains from generative AI: Experimental evidence

■ Purpose of Research

Recent studies have documented that the use of generative AI leads to productivity gains. However, in this strand of literature, there have not been many discussions on the mechanism through which generative AI can improve productivity. This study focuses on a division-of-labor-related pathway among possible generative-AI-driven productivity-improvement mechanisms to analyze how generative AI use and the division of labor interact and their implications on productivity gain by conducting an online experiment.

■ Main Outcomes of Research and Policy Implications

The experiment employed a 2×2 design: depending on whether generative AI is used and whether there is a division of labor, there are four groups. Each group completed a marketing-related task for a

hypothetical product. The marketing task consists of two subtasks that can be completed either through a division of labor or sequentially by a single individual.

First, we compare, under the absence of generative AI, the group that works on the task under a division of labor with the other group that does so without a division of labor. We find that the output quality from the former is lower than the latter, which reaffirms the inefficiency of the division of labor—previously discussed in the literature—due to information loss and the costs of interpreting conveyed information.

Second, to measure productivity changes in a situation where generative AI makes the division of labor unnecessary, we compared the group that completed the task with a division of labor but without generative AI with the group that completed the task without a division of labor while using generative AI. The result of this comparison demonstrates a productivity-improvement pathway, from the use of generative AI through the disappearance of the division of labor to higher productivity. In other words, when generative AI allows an individual to complete tasks that previously required division of labor among more than one person, the elimination of the negative effects on productivity due to the division of labor and direct benefit from using generative AI altogether improves productivity.

Finally, to examine how generative AI affects the division of labor, we compared two groups that completed the task with generative AI. Two groups differ in whether the task was completed under the division of labor or not. We find that there is no statistically significant

difference in output quality between two groups. This suggests that when generative AI is used under a setting with a division of labor, it may offset the inefficiencies from the division of labor. More broadly, these findings imply that productivity gains by generative AI that replaces people and discards the division of labor are not much different from productivity gains by generative AI that complements people under the division of labor.

정보통신정책연구원 기본연구 안내

■ 2023 기본연구

기본연구 23-01 데이터 경제 활성화를 위한 민관 역할분담과 정책 개입영역 연구(윤성욱, 박소연)

기본연구 23-02 해외 주요국 신규 사업자 이동통신시장 진입 효과 분석(김민희)

기본연구 23-03 유사국 사례를 통해 본 북한 통신시장 발전 방안 연구(임동민, 서소영, 이종화, 조대근, 서흥수)

기본연구 23-04 메타버스 시대 기본권 보호에 관한 연구: 인격권 논의를 중심으로(권은정, 한혜지, 계인국, 김법연, 이승민)

기본연구 23-05 디지털화폐 생태계 변화에 대응한 중앙은행 디지털화폐(CBDC) 도입 정책 연구(박동욱)

기본연구 23-06 ICT 확산에 따른 노동시장 임금격차(최지은, 이은영, 최세림, 이현옥)

기본연구 23-07 데이터 생산·관리 역량 평가를 위한 성숙도 모형 개발 연구(노희용, 장신재, 박지원)

기본연구 23-08 인터넷 생태계 환경 변화에 따른 네트워크 인프라 비용 분담에 관한 연구(염수현, 강인규, 이상규, 김태오)

기본연구 23-09 미디어 이용 조사의 모드 효과 비교(신지형)

기본연구 23-10 방송미디어분야 자율규제 제도화 방안 연구(이종원)

기본연구 23-11 이동통신 네트워크 인프라 산업 생태계의 전환 방향 연구(여재현, 박지현, 윤도원, 장희선, 김선우, 정인준, 박의환)

■ 2024 기본연구

기본연구 24-01 AI 반도체 생태계 조사(윤성욱, 김민식, 민대홍)

기본연구 24-02 인공지능 감시에 의한 권력의 확대와 그 규범적 대응방안 연구(문광진, 허진주, 전민경)

기본연구 24-03 시청각미디어 콘텐츠의 공정한 거래를 위한 정책 방안 연구: 수익 배분 관련 이슈를 중심으로(강준석)

기본연구 24-04 온라인 광고 시장의 경매 경쟁에 대한 연구(백소성)

기본연구 24-05 기술진보에 따른 산업별 업무 변화 연구(서영선)

기본연구 24-06 5G 요금제 특성이 이용자 편익에 미치는 영향: 단말기와 요금제의 결속을 중심으로(박진환)

기본연구 24-07 방송 프로그램 장르 구분의 타당성 검토: 방송법과 방송사업자 및 수용자의 장르 분류 기준 차이를 중심으로(황현정)
기본연구 24-08 생성형 AI가 미디어 분야에 미칠 영향에 대한 탐색적 연구(곽동균, 김남두, 주성희, 황현정, 강하연, 김예빈)
기본연구 24-09 디지털 경제 활성화를 위한 실물자산토큰(Real World Asset) 시장의 탈중앙화 금융 알고리즘 적용 가능성 연구(김 찬)
기본연구 24-10 AI 반도체 정책의 효과성 제고 방안 연구(정현준, 김민식, 오정숙)

■ 2025 기본연구

기본연구 25-01 통신시장의 경쟁활성화를 위한 공동지배력 평가 방법론 연구 (라성현)
기본연구 25-02 디지털 심화에 따른 몰입형 서비스 생태계 연구 (이재영, 심홍진, 성윤희, 이승환, 구자근)
기본연구 25-03 중국 유통플랫폼의 글로벌 확장과 대응방안 (김성옥, 장신재, 손가녕, 전민경)
기본연구 25-04 디지털 융합 시대 정보통신법제의 통합과 재편에 관한 연구 (권은정, 김법연, 김지훈, 선지원, 박상미)
기본연구 25-05 디지털 환경 변화를 반영한 데이터 트래픽 예측 연구(김민희, 정광재)
기본연구 25-06 글로벌 미디어 플랫폼 확산 대응을 위한 국내 미디어 서비스 경쟁력 분석 및 강화 전략 연구(강준석, 권용재)
기본연구 25-07 월드뱅크 등 국제기구와의 협력 연구 (황준호, 양종민, 김지원, 문용일, 이효원)
기본연구 25-08 생성형 AI의 생산성 분석(민대홍)
기본연구 25-09 디지털 전환에 따른 소매패턴 변화와 정책 방향(장신재, 이선희)
기본연구 25-10 인공지능·자율시스템 기반 도시의 윤리적 설계방안 연구 (이현경, 문광진, 전민경)
기본연구 25-11 디지털 기술 활용이 기업 성과에 미치는 영향 분석과 정책 시사점 연구(손녕선)
기본연구 25-12 생성형 인공지능 서비스 채택 선행요인에 관한 탐색적 연구 (주성희, 임연수, 김예빈)
기본연구 25-13 주파수 경매에서 네트워크 구축조건 도입 방식에 대한 연구(백소성)
기본연구 25-14 AI 기술주권과 국가경쟁력 제고 방안 연구

(이경선, 김성옥, 이경은, 오장민, 윤혜선)

기본연구 25-15 플랫폼 확산의 인플레이션 영향에 대한 연구(김경은, 조수진, 심명규)

기본연구 25-16 AI 시대의 포용과 상생을 위한 사회적 의제 연구
(문아람, 문정욱, 조성은, 연소라, 김휘홍, 이으뜸, 전민경, 신진호)

기본연구 25-17 ICT 혁신이 사회·경제에 미치는 영향(최지은, 서영선, 노희용)

기본연구 25-18 차세대 통신기술이 통신시장 경쟁구조 및 네트워크 투자에 미치는 영향 분석(김민희, 김경모)

● 저 자 소 개 ●

민 대 홍

- 성균관대 경제학과 졸업
- 성균관대 경제학 석사
- University of Arizona, Tucson 경제학 박사
- 현 정보통신정책연구원 연구위원

기본연구 25-08

생성형 AI의 생산성 분석

2025년 12월 일 인쇄
2025년 12월 일 발행

발행인 이 상 규
발행처 정 보 통 신 정 책 연 구 원
충청북도 진천군 덕산읍 정통로 18
TEL: 043-531-4114 FAX: 043-535-4695～6
인 쇄 인 성 문 화
ISBN 979-11-7000-429-5 93320

〈비매품〉