creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

박사학위논문

중소기업 맞춤형 ESG 컨설팅을 위한 LLM 기반 Agentic AI 프레임워크

HANSUNG UNIVERSITY

2026년

한 성 대 학 교 대 학 원

스마트융합컨설팅학과

스 마 트 융 합 제 품 전 공

노 은 미

HANSUNG
UNIVERSITY

박사학위논문
지도교수 노광현

중소기업 맞춤형 ESG 컨설팅을 위한 LLM 기반 Agentic AI 프레임워크

An LLM-Based Agentic AI Framework for Tailored ESG Consulting in Small and Medium-sized Enterprises

HANSUNG UNIVERSITY

2026년 6월 일

한 성 대 학 교 대 학 원

스마트융합컨설팅학과

스마트융합제품전공

노 은 미

박 사 학 위 논 문
지도교수 노광현

중소기업 맞춤형 ESG 컨설팅을 위한 LLM 기반 Agentic AI 프레임워크

An LLM-Based Agentic AI Framework for Tailored ESG Consulting in Small and Medium-sized Enterprises

위 논문을 공학 박사학위 논문으로 제출함

2026년 6월 일

한 성 대 학 교 대 학 원

스마트융합컨설팅학과

스마트융합제품전공

노 은 미

노은미의 공학 박사학위 논문을 인준함

2026년 6월 일

심사위원장 이 후 진 (인)

심 사 위 원 김 승 천 (인)

심 사 위 원 남 현 우 (인)

심 사 위 원 임 황 용 (인)

심 사 위 원 노 광 현 (인)

HANSUNG UNIVERSITY

국 문 초 록

중소기업 맞춤형 ESG 컨설팅을 위한 LLM 기반 Agentic AI 프레임워크

한 성 대 학 교 대 학 원
스 마 트 융 합 컨 설 팅 학 과
스 마 트 융 합 제 품 전 공
노 은 미

ESG(환경·사회·지배구조) 경영은 글로벌 공급망의 필수 요건으로 자리매김하였으며, 2024년부터 시행 중인 EU 기업 지속가능성 보고 지침과 국제지속가능성기준위원회 공시 기준은 대기업뿐 아니라 공급망에 포함된 중소기업에도 ESG 정보 공개 압력을 확대하고 있다. 그러나 국내 중소기업의 ESG 경영은 비용과 전문 인력 부담이라는 구조적 장벽에 직면해 있으며, 전문 ESG 컨설팅은 표준 절차상 상당한 비용과 시간을 요구한다. 본 연구는 이러한 비용·시간 장벽을 해소하면서도 전문가 수준의 ESG 컨설팅 품질을 달성하기 위해, 복수의 AI Agent가 체계적으로 협업하는 Agentic AI 기반 중소기업 ESG 맞춤형 LLM 컨설팅 시스템을 설계·구현·검증하였다.

본 연구의 방법론적 구조는 중소기업 ESG 컨설팅의 전체 과정을 인식(Perceive), 추론(Reason), 계획(Plan), 실행(Act), 학습(Learn)의 5단계로 구조화한 Agentic AI 프레임워크이다. 이 프레임워크는 단순한 순차 처리가 아니라, 각 단계가 다음 단계의 입력 맥락을 구조적으로 제공하고 필요 시 이전

단계로 피드백을 환류하는 폐루프(closed-loop)로 작동한다는 점에 핵심이 있다. 상위 오케스트레이터(Orchestrator)가 각 단계를 조율하여 자가진단 입력에서 등급 예측, 맞춤형 컨설팅 생성, 전문가 검증으로 이어지는 흐름을 하나의 파이프라인으로 통합하며, 본 연구는 이 가운데 자가진단 입력부터 컨설팅 생성까지의 Stage 1~3을 구현·검증 범위로, 전문가 검증에 해당하는 Stage 4를 개념 실증 범위로 설정하였다.

Agentic AI 프레임워크는 상위 오케스트레이터의 조율 아래 세 개의 협력 Agent로 구체화된다. ESG 등급 예측 Agent는 중소기업의 자가진단 응답을 입력받아 Gradient Boosting 기반 3클래스 분류 모델로 처리하여, 예측 등급, 클래스별 확률분포, SHAP 기반 취약 영역 설명의 다차원 정보를 출력한다. LLM 기반 ESG 컨설팅 Agent는 이 진단 결과를 입력 맥락으로 받아 RAG(검색증강생성)와 지식그래프(KG)의 듀얼 채널로 관련 법규, 인증, 모범사례를 검색·연결하여 맥락을 강화하고, 진단결과 요약, 맞춤형 컨설팅 보고서, 증빙 서식 가이드의 세 가지 맞춤형 산출물을 생성한다. 전문가 검증 Agent는 생성된 산출물을 입력으로 받아 EIEC(ESG Improvement Execution Contribution) 5차원 품질 프레임워크에 따라 전문가가 독립적으로 평가하는 HITL(Human in the Loop) 구조로 처리하며, 차원별 품질 점수, 평가자 간 일치도, 강점 및 우려를 출력하여 앞선 두 Agent의 개선 근거로 환류한다. 세 Agent는 독립 모듈의 단순 결합이 아니라, 예측 결과가 컨설팅의 입력으로, 컨설팅 산출물이 다시 검증의 대상으로 전달되는 하나의 유기적 파이프라인으로 연결된다.

이상의 시스템 아키텍처는 실제 동작하는 시스템으로 구현되었다. 본 연구는 세 협력 Agent와 상위 오케스트레이터, RAG와 KG의 듀얼 채널, 그리고 자가진단 웹 인터페이스를 하나로 통합하여, 기업 담당자의 자가진단 응답 입력에서 등급 예측, 맥락 강화, 맞춤형 산출물 생성에 이르는 전 과정을 별도의 설치 없이 사용할 수 있는 단일 온라인 서비스(플랫폼)로 구현하였다. 이를 통해 개념적 설계가 자가진단 입력부터 컨설팅 산출물까지 작동하는 실증 가능한 시스템으로 구체화되었다.

이렇게 구현한 시스템을 비상장 중소기업 299개사를 대상으로 맥락 조건

과 모델의 교차 설계로 실증한 결과, 자가진단 자기보고 응답에 파생 정량신호를 결합한 조건에서 ESG 등급 예측 정확도(Accuracy) 0.8631을 달성하여 자가진단 데이터만으로 ESG 평가 등급을 예측할 수 있음을 확인하였고, RAG와 KG를 결합한 맥락 강화 조건에서 컨설팅 품질 총점 8.52점으로 최고 품질을 달성하여 맥락 강화의 효과를 검증하였다. 이 밖에 토큰 효율성에 기반한 이원적 배포 전략, 산업적합성의 업종 간 차이, AI 재현성의 일관성도 함께 확인하였다. 나아가 산출물을 실제로 사용하는 수요기업을 대상으로 한 만족도 검증에서 전반적 만족도는 5점 척도 평균 4.25점(만족 이상 83.3%), 도입 의향은 평균 4.08점으로 나타나, 전문가 품질 검증과는 별개로 실무 수용성 측면에서도 산출물의 활용 가능성이 확인되었다.

본 연구는 비상장 중소기업의 자가진단 데이터만으로 ESG 평가 등급 수준으로 예측이 가능함을 실증하고, ESG 자가진단과 AI 컨설팅을 단일 Agentic AI에서 멀티 Agent 파이프라인으로 통합하면서 EIEC 5차원 품질평가 프레임워크와 ESG 브릿지 도구(ESG Bridge Tool) 개념을 제안·검증하였다. 또한 이를 전문가 수준의 컨설팅을 저비용·단시간에 제공하는 온라인 서비스로 구현하고, HITL 평가를 통해 AI 자동화가 적절한 영역과 전문가 개입이 필요한 영역을 구분하여 실무 배포의 근거를 제시하였다. 향후 Stage 4·5의 본격 구현과 종단적 관측을 통한 자율적 개선 구조로의 확장이 후속 과제로 남는다.

【주제어】 ESG 컨설팅, 중소기업, Agentic AI, 거대언어모델(LLM), 지식그래프, 전문가 검증

목 차

제 1 장 서 론 ····· 1

제 1 절 연구 배경 및 필요성 ····· 1

제 2 절 연구 목적 및 내용 ····· 5

제 2 장 이론적 배경 및 선행연구 ····· 7

제 1 절 ESG 경영과 중소기업 ····· 7

제 2 절 ESG 평가 체계 및 자가진단 ····· 8

제 3 절 머신러닝 기반 ESG 예측 연구 ····· 10

제 4 절 LLM과 ESG 컨설팅 자동화 ····· 11

제 5 절 Agentic AI의 개념과 구조 및 ESG 적용 ····· 14

제 3 장 Agentic AI 기반 ESG 컨설팅 프레임워크 ····· 21

제 1 절 ESG 컨설팅 프레임워크 설계 ····· 21

제 2 절 LLM 기반 Agentic AI 구성 및 설계 ····· 23

제 3 절 자가진단 문항 설계 및 중요도 분석 ····· 26

제 4 장 ESG 등급 예측 Agent ····· 29

제 1 절 ESG 등급 예측 Agent 설계 ····· 29

제 2 절 ESG 등급 예측 Agent 실험결과 ····· 36

제 3 절 ESG 등급 예측 성능 종합 및 컨설팅 연계 ····· 45

제 5 장 LLM 기반 ESG 컨설팅 Agent ····· 47

제 1 절 LLM 기반 ESG 컨설팅 Agent 설계 ····· 47

제 2 절 LLM 컨설팅 산출물 ····· 52

제 3 절 실험 설계: 3×2 완전교차 ····· 53

제 4 절 EIEC 5차원 품질평가 통합 프레임워크 ····· 60

제 5 절 컨설팅 Agent 설계 종합 ····· 70

제 6 장 전문가 검증 Agent ····· 72
제 1 절 전문가 검증 Agent 설계 ····· 72
제 2 절 전문가 검증 Agent 실험 결과 ····· 74

제 7 장 프레임워크 구현 및 성능 분석 ····· 89
제 1 절 맥락 강화의 품질 향상 효과 ····· 89
제 2 절 품질-연산 효율성 분석 ····· 97
제 3 절 업종별 컨설팅 품질 차별성 분석 ····· 99
제 4 절 AI 재현성 검증 ····· 106
제 5 절 ESG 자가진단 온라인 서비스 설계 및 구현 ····· 107
제 6 절 ESG 수요기업 만족도 검증 ····· 125
제 7 절 실증 결과 종합 정량 분석 ····· 128

제 8 장 결 론 ····· 131
제 1 절 연구요약 ····· 131
제 2 절 학술적 및 실무적 기여도 ····· 133
제 3 절 향후 연구 방향 ····· 135

참 고 문 헌 ····· 137

별 첨 문 서 ····· 146

ABSTRACT ····· 150

HANSUNG UNIVERSITY

표 목 차

[표 2-1] ESG 평가기관·자가진단 도구 유형별 비교 ····· 9
[표 2-2] ML 기반 ESG 등급 예측 선행연구 비교 ····· 11
[표 2-3] LLM 기반 ESG·RAG·KG 선행연구 요약 ····· 13
[표 2-4] 선행연구 대비 본 연구의 차별성 매핑 ····· 20
[표 3-1] 5단계 ESG 컨설팅 프레임워크 ····· 21
[표 3-2] 연구 내용 및 분석 방법 체계 ····· 25
[표 4-1] GBM 하이퍼파라미터 탐색 범위 및 최종 설정 ····· 30
[표 4-2] 대안 모델 비교 (조건 A, Stratified 5-Fold CV) ····· 31
[표 4-3] ESG 자가진단 30문항 및 대응 증빙 양식 ····· 32
[표 4-4] 종속변수(ESG 등급) 3클래스 분포 ····· 34
[표 4-5] 피처 조건 설계 (조건 A·B·C·C-1) ····· 35
[표 4-6] 조건별 예측 성능 (GBM, Stratified 5-Fold CV) ····· 36
[표 4-7] Feature Importance 상위 10개 변수 (GBM, 조건 C-1) ····· 39
[표 4-8] SHAP Feature Importance 상위 10개 (XGBoost, 조건 B) ···· 41
[표 4-9] 산업군별 ESG 자가진단 기술통계 ····· 44
[표 4-10] Stratified 5-Fold 클래스 분포 ····· 46
[표 5-1] RAG 지식 베이스 구성 ····· 49
[표 5-2] LLM 컨설팅 산출물 구성 ····· 52
[표 5-3] 3×2 완전교차 실험 설계 ····· 54
[표 5-4] 자가진단 플랫폼(esg-platform-v8) 핵심 함수와 역할 ····· 57
[표 5-5] 관리자 콘솔(esg-admin-delta) 핵심 함수와 역할 ····· 59
[표 5-6] EIEC 5차원 정의 및 측정 기준 (5점 리커트 척도) ····· 64
[표 5-7] EIEC 측정모형 타당성 검증 결과 ····· 69
[표 6-1] 전문가 평가 결과 ····· 76
[표 6-2] 전문가 평가 측정도구 신뢰도 및 평가자 간 일치도 (n=15) ···· 77
[표 6-3] 전문가 패널 인구통계학적 구성 (n=15) ····· 79
[표 7-1] 조건 × 모델별 LLM 컨설팅 품질 총점 ····· 89
[표 7-2] 맥락 조건 간 ANOVA 및 Tukey HSD 사후검정 결과 ····· 91

[표 7-3] 조건 × 모델별 토큰 사용량 및 효율성 ····· 98
[표 7-4] 업종별 Baseline vs RAG+KG 품질 총점 (Sonnet) ····· 101
[표 7-5] AI 재현성 검증 결과 ····· 106
[표 7-6] 국내외 ESG 자가진단 온라인 서비스 비교 ····· 110
[표 7-7] 시스템 스택 요약 (3-시스템 분리 아키텍처) ····· 111
[표 7-8] 전문 ESG 컨설팅 대비 자동화 시스템 4축 비교 ····· 123
[표 7-9] 연구 영역별 주요 발견 및 기여 요약 ····· 124
[표 7-10] 수요기업 만족도 검증 결과 (n=12, 5점 척도) ····· 126
[표 7-11] 단일 LLM 호출 대비 Agentic 파이프라인 구조 효과 ····· 129

HANSUNG UNIVERSITY

그 림 목 차

[그림 1-1] 연구범위 — ESG 평가 생태계에서의 위치 ····· 3
[그림 2-1] Agentic AI 워크플로 8대 패턴과 본 연구의 적용 범위 ····· 17
[그림 2-2] 본 연구 파이프라인의 Agentic AI 패턴 매핑 ····· 18
[그림 3-1] Agentic AI 5단계 프레임워크 구조도 ····· 24
[그림 4-1] ESG 등급 예측 Agent의 구조와 동작 흐름 ····· 29
[그림 4-2] 조건별 예측 성능 비교 (Accuracy, F1-Score) ····· 38
[그림 4-3] Feature Importance 분석 결과 (GBM, 조건 B) ····· 40
[그림 4-4] SHAP Feature Importance (XGBoost, 조건 B) ····· 42
[그림 4-5] 환경·사회·지배구조 영역별 SHAP 기여도 ····· 43
[그림 5-1] LLM 기반 ESG 컨설팅 Agent의 구조와 동작 흐름 ····· 47
[그림 5-2] EIEC 3층 계층 구조 ····· 60
[그림 5-3] EIEC 3대 이론적 기반 매핑 ····· 61
[그림 5-4] EIEC 5차원 확인적 요인분석 요인부하량 (n=30, 예비) ···· 67
[그림 6-1] 전문가 검증 Agent의 구조와 동작 흐름 ····· 72
[그림 6-2] 전문가 평가 설문지 v4.0 5섹션 구조 ····· 74
[그림 6-3] 토큰 기반 익명 응답 흐름 ····· 75
[그림 6-4] 5개 평가항목 평균 점수 (n=15) ····· 78
[그림 6-5] 응답자별 총점 분포 (5개 항목 합산, n=15) ····· 79
[그림 6-6] 응답자별 × 평가항목별 점수 히트맵 (1~5점, n=15) ····· 80
[그림 6-7] 5개 평가항목 레이더 차트 (평균·최고·최저) ····· 81
[그림 6-8] 전문가 대비 AI 품질 수준 평가 (C1, n=15) ····· 82
[그림 6-9] 강점 vs 우려사항 매트릭스 (C4 vs C5, n=15) ····· 83
[그림 6-10] HITL 피드백 신호의 Agent 파이프라인 매핑 ····· 85
[그림 6-11] HITL 필수 영역 vs AI 강점 영역 비교 ····· 86
[그림 6-12] Agent별 개선 우선순위 매트릭스 ····· 87
[그림 7-1] 조건별 5차원 품질 프로파일 (Sonnet 모델 기준) ····· 92
[그림 7-2] 모델별 5차원 품질 프로파일 (RAG+KG 조건 기준) ····· 94
[그림 7-3] 업종별 Baseline vs RAG+KG 품질 총점 (Sonnet, n=30) ·· 100

[그림 7-4] 업종별 5차원 품질 프로파일 (RAG+KG, Sonnet 모델) …… 103
[그림 7-5] 3-시스템 통합 아키텍처 …… 111
[그림 7-6] esg-platform-v8 자가진단 시작 화면 …… 113
[그림 7-7] esg-platform-v8 자가진단 30문항 응답 화면 …… 115
[그림 7-8] esg-agent-admin 토큰 발급·응답 모니터링 화면 …… 116
[그림 7-9] esg-survey 전문가 응답 화면 …… 117
[그림 7-10] 자가진단 보고서 PDF 첫 페이지 (테스트 기업 v18) …… 118
[그림 7-11] 컨설팅 보고서 PDF 핵심 페이지 (테스트 기업 v18) …… 119
[그림 7-12] 증빙 서식 워드 파일 5종 대표 페이지 …… 120
[그림 7-13] Supabase 데이터베이스 스키마 다이어그램 …… 122
[그림 7-14] 수요기업 주요 평정 지표(만족·도입·추천) 평균 (n=12) …… 126
[그림 7-15] 수요기업이 평가한 보고서 품질 5차원 평균 (n=12) …… 128

HANSUNG UNIVERSITY

제 1 장 서 론

제 1 절 연구 배경 및 필요성

1. 연구배경

ESG(Environmental, Social, Governance) 경영은 20세기 후반의 기업 사회적 책임(Corporate Social Responsibility, CSR) 논의에서 출발하여, 2000년대 들어 UN 책임투자원칙[1] 과 지속가능성 공시 기준의 국제화를 거치면서 오늘날 기업의 비재무적 성과를 평가하는 핵심 프레임워크로 자리매김하였다. 특히 2010년대 중반 이후 글로벌 자산운용사와 연기금을 중심으로 ESG 투자 자산이 급증하면서, ESG는 선택적 윤리 경영 단계에서 벗어나 자본시장 접근성을 결정하는 필수 요건 단계로 진화하였다. 이러한 변화는 EU 비재무정보 공시지침[2]과 지속가능성 보고 지침, 그리고 ISSB의 S1·S2 공시 기준(2023)으로 제도화[3]되었으며, 대기업의 ESG 공시 의무가 공급망을 통해 중소기업에 파급되는 구조를 형성하였다.

McKinsey & Company[4]의 글로벌 설문조사에 따르면 전 세계 기업의 71%가 생성형 AI를 업무에 활용하고 있으나, AI 도입을 통해 전사적 EBIT(이자·세전 이익) 차원의 실질적 기여를 확보했다고 보고한 기업은 아직 소수에 그친다. 이는 AI 기술의 확산 속도(adoption rate)와 실질 비즈니스 가치 창출 속도(value realization rate) 사이에 상당한 격차가 존재함을 의미하며, 학술·산업 문헌에서는 이를 Lab-to-Market 갭이라 명명한다. 이 용어는 연

---

1) Principles for Responsible Investment (PRI) (2006). Principles for Responsible Investment. UNEP Finance Initiative & UN Global Compact.

2) European Union (2014). Directive 2014/95/EU (Non-Financial Reporting Directive). Official Journal of the European Union.

3) 국제지속가능성기준위원회(ISSB, International Sustainability Standards Board): IFRS 재단이 2021년 설립한 글로벌 지속가능성 공시기준 제정 기구로, 2023년IFRS S1·S2 기준 발표함.

4) Singla, A., Sukharevsky, A., Yee, L., Chui, M., & Hall, B. (2025). The state of AI: How organizations are rewiring to capture value. McKinsey & Company.

구실(lab) 단계에서 검증된 AI 기술이 실제 시장(market)에서 운영 가능한 서비스로 안착하기까지의 성과·접근성 격차를 의미한다. 본 연구의 문제의식은 이 Lab-to-Market 갭이 ESG 컨설팅 영역에서 특히 첨예하게 나타난다는 관찰에서 출발한다. LLM과 검색증강생성(RAG)[5]를 활용한 ESG 연구는 2023년 이후 급격히 증가하였지만, 그 대부분은 대기업 공시 문서의 품질 평가나 ESG 점수의 텍스트 기반 추정에 초점을 두고 있을 뿐, 중소기업이 실제로 접근 가능한 운영 가능한 서비스 수준으로 구현된 사례는 드물다.

국내 맥락에서는 산업통상자원부가 2021년 K-ESG 가이드라인 v1.0을, 2022년 v2.0을 발표하였으며, 2024년 12월 개정판을 추가로 공개하였다. 또한 중소벤처기업부·중소기업중앙회는 중소기업에 특화된 자가진단 체계를 운영해 왔다. 그러나 자가진단 결과의 해석, 구체적 개선과제 도출, 관련 법규·인증 기준과의 매핑과 같은 진단 이후의 단계는 여전히 전문 컨설턴트의 개입에 의존하고 있다. 중소벤처기업진흥공단의 「中企 ESG 성과분석을 통한 중소기업 지원방안 연구」에 따르면, 조사 대상 중소기업 300개사 중 37.0%가 ESG 경영 도입의 최대 애로사항으로 비용 부담을 지목하였다. 실제 컨설팅 비용을 보면, 중소기업중앙회의 ESG 컨설팅 사업(2021)은 일반기업 부담 비용을 1건당 200만 원으로 책정하였다. 또한 사전교육, 온라인 자가진단, 현장실사, 보고서 배부에 이르는 표준 수행절차는 수일에서 수주의 기간을 요한다. 이러한 비용·시간 부담은 연매출 100억 원 미만의 중소기업에 상당한 재무·시간 제약으로 작용하며, 그 결과 자가진단을 완료한 기업 중 실제 후속 개선 활동까지 이어지는 비율은 제한적이어서 진단과 실행 사이의 괴리(planning-action gap) 현상으로 관찰된다.

이러한 구조적 장벽을 해소하기 위해서는 단일 LLM 호출로 개선안을 뽑아내는 접근을 넘어, 복수의 AI Agent가 각기 다른 역할을 분담하여 체계적으로 협업하는 구조가 요구된다. Agentic AI는 자율적 인식(Perceive)에서 추론(Reason), 계획(Plan), 실행(Act)을 거쳐 학습(Learn)에 이르는 반복 사이클을 수행하는 다중 Agent 시스템으로[6], 기존 단일 LLM 호출 방식이 갖는 환

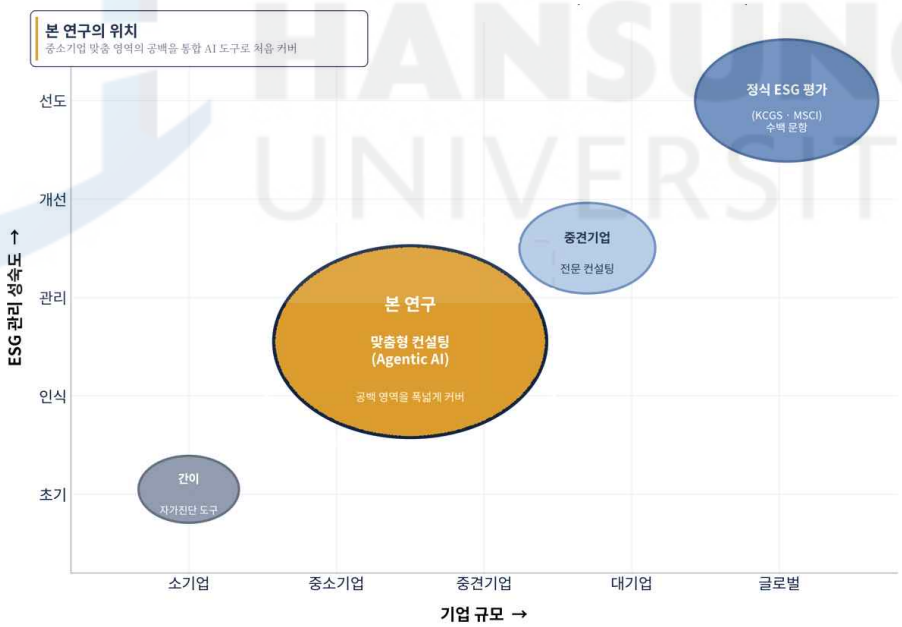
5) 검색증강생성(RAG, Retrieval-Augmented Generation): 외부 지식 검색 결과를 LLM의 입력 맥락으로 결합하여 환각을 줄이고 사실성을 높이는 생성 기법.

6) Acharya, D. B., Kuppan, K., & Divya, B. (2025). Agentic AI: Autonomous intelligence for

각(hallucination), 맥락 부족, 일관성 결여 문제를 구조적으로 완화할 수 있는 유력한 대안으로 부상하고 있다. 본 연구는 이 Agentic AI 프레임워크를 중소기업 ESG 컨설팅 영역에 특화하여 재설계함으로써, 비용·시간 장벽을 해소하면서도 전문가 수준의 품질을 유지하는 실제 배포 가능한 시스템을 설계·구현·검증한다.

2. 연구의 필요성

본 연구의 필요성을 명확히 하기 위하여 현재의 ESG 평가 생태계를 기업 규모(x축)와 ESG 성숙도(y축)의 2차원 좌표 평면에 배치하고 주요 도구와 제도의 위치를 점검하였다. 이 지형도는 기존 ESG 연구가 다루어 온 대상과 그 대상들 사이의 빈 공간(research gap)을 시각적으로 드러낸다.



[그림 1-1] 연구 범위 — ESG 평가 생태계에서의 위치

complex goals – A comprehensive survey. IEEE Access, 13, 18912–18936. ; Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., et al. (2024). A survey on large language model based autonomous agents. Frontiers of Computer Science, 18(6), Article 186345.

좌측 하단의 저규모·저성숙도 영역에는 중소벤처기업진흥공단 ESG 자가진단, K-ESG 가이드라인 자가진단 툴, 산업통상자원부 ESG 진단 툴 등이 위치한다. 이 영역의 도구들은 진입 장벽이 낮고 사용이 간편하다는 장점이 있으나, 진단 결과의 해석과 구체적 실행 방안 도출은 사용자에게 맡겨지는 한계를 공유한다. 우측 상단의 대규모·고성숙도 영역에는 MSCI ESG, 한국ESG기준원 ESG 등급, SASB, S&P Global ESG, GRI Standards, Bloomberg ESG 등 정식 ESG 평가 체계가 배치된다. 이 영역은 상장 대기업을 전제로 설계된 정교한 평가 체계이지만, 중소기업의 경영 현실과 자원 제약을 직접 반영하기에는 부적합하다. 중앙의 중규모·중간 성숙도 영역에는 대형 컨설팅 펌의 ESG 자문, 업종별 ESG 평가 기관, ESG 인증기관이 활동하고 있다. 이 영역은 고비용의 전문 자문 형태로, 재무 여력이 있는 중견기업에 주로 접근 가능하다. 결과적으로 이 지형에서 체계적으로 비어 있는 영역은, 중소기업이 정식 ESG 평가 응시 이전의 준비 단계에서 자가진단과 맞춤형 개선 지원을 통합적으로 활용할 수 있는 도구의 부재이다.

본 연구는 이러한 공백을 해소하는 것을 목표로 한다. 즉 본 연구는 자가진단과 맞춤형 개선 지원을 AI 기반으로 통합 제공하는 도구를 Agentic AI 프레임워크로 설계 및 구현하는 연구로 정의되며, 이러한 도구를 ESG 브릿지 도구(ESG Bridge Tool)로 명명한다. 브릿지라는 명칭은 Porter[7]가 경영전략론에서 사용한 전환적 도구(transitional tool) 개념에서 유래한 것으로, 한 상태에서 다음 상태로의 성숙 단계 이행을 지원하는 도구를 의미한다. 본 연구의 대상은 중소기업이 K-ESG 자가진단을 넘어 MSCI·한국ESG기준원 수준의 정식 평가에 대응 가능한 준비도에 도달할 수 있도록, 자가진단 결과를 예측·해석하고 맞춤형 개선과제를 자동 생성하며 증빙 서식까지 제공하는 통합 AI 시스템이다.

7) Porter, M. E. (1996). What is strategy? Harvard Business Review, 74(6), 61-78.

제 2 절 연구 목적 및 내용

본 연구의 목적은 Agentic AI 기반 멀티 Agent 파이프라인을 하나의 Agentic AI 프레임워크로 설계 및 구현하여, 중소기업 ESG 맞춤형 컨설팅 방법의 기술적 타당성과 실무 적용 가능성을 실증 데이터와 전문가 평가를 통해 검증하는 데 있다. 이 목적은 세 가지 세부 목표로 구체화된다. 첫째, 중소기업 자가진단 문항 응답만으로 외부에서 독립적으로 산정된 ESG 등급을 예측할 수 있는 머신러닝 모델을 개발하고, 산업분류, E/S/G 자가진단 평균, 파생 정량신호의 단계적 추가가 예측 성능을 어떻게 변화시키는지를 체계적으로 분석한다. 둘째, RAG와 지식그래프(KG)[8]를 결합한 듀얼 채널 맥락 강화가 LLM이 생성하는 ESG 컨설팅 보고서의 품질을 유의하게 개선하는지를 3×2 완전교차 실험 설계로 검증한다. 셋째, 추가 맥락 투입에 따른 품질 향상과 연산 자원(토큰) 소비의 관계를 정량화하여, 실무 배포 시 비용·품질 트레이드오프를 명시적으로 고려한 이원적 배포 전략의 근거를 제시한다.

이 세 가지 목적에 대응하여 본 연구의 연구 내용은 ESG 등급 예측, 맥락 강화 품질, 품질·연산 효율성의 세 가지로 구분된다. ESG 등급 예측은 중소기업 자가진단 30문항 응답으로 ESG 등급(3클래스: Low/Mid/High)을 예측하고, 산업분류, E/S/G 자가진단 평균, 파생 정량신호의 단계적 추가가 예측 성능에 미치는 영향을 분석한다. 맥락 강화 품질은 RAG와 KG 기반 맥락 강화가 LLM 컨설팅 품질(구체성, 실행가능성, 산업적합성, 구조화, 추적가능성의 5차원)을 개선하는지를 3개 맥락 조건(Baseline/RAG/RAG+KG)과 2개 모델(Claude Haiku/Sonnet)의 완전교차 설계로 검증하고, 조건·모델·상호작용 효과를 분석한다. 품질·연산 효율성은 추가 맥락 투입(토큰 증가)과 품질 향상의 관계를 규명하고, 품질·연산 효율성(품질/1K 토큰) 관점에서 실무 배포 시 최적 조건을 도출한다.

세 연구 내용은 Agentic AI 프레임워크 관점에서 상호 연결되며, 독립된 세 실험의 나열이 아니라 파이프라인의 각 단계가 서로를 강화하는지를 통합적으로 검증하는 설계이다.

---

8) 지식그래프(KG, Knowledge Graph): 개체와 관계를 노드와 엣지로 구조화한 지식 표현으로, 본 연구에서는 ESG 규제·인증·업종 관계를 연결한다.

구체적으로 본 연구가 제시하는 전체 방법은 Agentic AI 5단계 프레임워크 인식(Perceive), 추론(Reason), 계획(Plan), 실행(Act), 학습(Learn)을 토대로 하며, 그 가운데 자가진단 입력부터 컨설팅 생성까지의 Stage 1~3을 구현·검증 범위로, HITL(Human in the Loop) 검증 단계인 Stage 4를 개념 실증 범위로 설정한다. 이 프레임워크는 세 개의 서브 Agent로 구체화된다. ESG 등급 예측 Agent는 자가진단 응답으로부터 ESG 등급과 취약 영역을 도출하여 ESG 등급 예측 분석에 대응하고, LLM 기반 ESG 컨설팅 Agent는 그 진단 결과를 입력받아 RAG와 KG로 맥락을 강화한 맞춤형 컨설팅 보고서를 생성하여 맥락 강화 품질 분석에 대응하며, 전문가 검증 Agent는 생성된 보고서를 EIEC 5차원과 전문가 평가로 검증한다. 품질·연산 효율성 분석은 이 파이프라인 전반의 품질·비용 트레이드오프를 정량화하여 실무 배포 전략의 근거를 마련한다. 따라서 본 연구의 목적은 개별 모델의 성능 입증을 넘어, 세 개의 Agent가 하나의 유기적 파이프라인으로 작동하는 통합 방법을 설계하고 그 타당성을 실증하는 데 있다.

HANSUNG UNIVERSITY

제 2 장 이론적 배경 및 선행연구

제 1 절 ESG 경영과 중소기업

ESG는 기업의 환경적 책임, 사회적 가치, 지배구조 투명성을 통합 평가하는 프레임워크이다. 2004년 유엔 글로벌 콤팩트 보고서에서 용어가 공식화된 뒤 2006년 UN 책임투자원칙(PRI)으로 국제 자본시장에 확산되었고, 2010년대 GRI, SASB, TCFD의 공시 기준화를 거쳐 2020년대에 제도화 및 의무화 단계로 진입하였다. 제도화의 분기점은 EU의 기업 지속가능성 보고지침(CSRD)과 공급망 실사지침(CSDDD)이며, IFRS는 S1, S2(2023)가 글로벌 공통 기준을 제시하였다. 이러한 규범은 공급망 경로(원청 대기업의 실사 대응에 따른 협력사 정보 요구)와 국내 제도화 경로(금융위원회의 단계적 의무화 로드맵)를 통해 한국 중소기업에 파급된다. 한국 중소기업 대다수는 당장 직접적 공시·실사 의무를 지지 않으나, 대기업·글로벌 원청의 대응이 본격화될수록 공급망을 통한 ESG 정보 요구가 불가피하게 확산되는 구조에 놓인다.

국내에서는 산업통상자원부가 K-ESG 가이드라인(2021 v1.0, 2022 v2.0, 2024년 개정)을, 중소벤처기업부, 중소기업중앙회가 중소기업 특화 간소화 자가진단 체계를 운영해 왔다. 상당수 중소기업은 ESG의 필요성은 인식하나 체계적 관리 역량은 부족한 것으로 보고되며, 전문 인력, 예산, 이해 부족이 주원인으로 컨설팅 접근성 확보가 정책 과제로 부각되어 왔다. 한편, 중소벤처기업진흥공단 조사에서 중소기업의 37.0%가 ESG 도입의 최대 애로사항으로 비용 부담을 꼽았으며, 컨설팅 단가는 1건당 200만 원 수준의 비용과 수일~수주의 소요 기간이 연매출 100억 원 미만 중소기업에 부담으로 작용한다.

국제 ESG 평가·공시 생태계는 상호 보완적 표준과 평가기관으로 구성된다. GRI(보고 원칙·형식의 글로벌 표준), SASB(77개 산업별 재무적 중요성 지표, 2022년 ISSB 통합), TCFD(기후 재무 공시 4대 영역, IFRS S2의 기초)가 대표적이며, 이 세 표준은 중소기업의 직접 적용 대상은 아니지만 대기업이 공급망 중소기업에 요구하는 ESG 정보의 범위와 형식을 결정한다.

정식 ESG 평가 체계는 평가기관에 따라 방법론과 점수가 상이하다. MSCI(AAA~CCC), Sustainalytics(리스크 점수), Bloomberg(정량 스코어), 국내 한국ESG기준원(S~D)·서스틴베스트 등이 각기 다른 척도를 사용하며, 평가기관별 점수의 상관관계는 0.3~0.7 수준으로 ESG 평가의 정답이 단일하지 않음이 학술적으로 확인된다. 본 연구는 이 중 하나를 채택하는 대신 외부에서 독립적으로 산정된 ESG 등급을 종속변수로 사용함으로써, 자가진단(자기보고)과 외부 평가(외부 관찰)의 완전 독립을 확보하면서 한국 중소기업의 실무 맥락에 근접한 기준을 종속변수로 삼는다.

제 2 절 ESG 평가 체계 및 자가진단

K-ESG 가이드라인은 정보공시(P) 5문항, 환경(E) 17문항, 사회(S) 22문항, 지배구조(G) 17문항의 총 61개 진단항목으로 구성되어 있다. 각 항목은 점검기준, 성과점검 방법, 배점(100점 만점), 추가설명을 포함하여 기업이 자율적으로 ESG 경영 수준을 진단할 수 있도록 설계되었다. 그러나 61개 항목을 모두 응답하고 해석하는 것은 중소기업의 현실적 역량을 초과하는 부담이며, 이에 중소벤처기업부·중소기업중앙회는 K-ESG 가이드라인의 핵심 항목을 추출하여 환경 10문항, 사회 10문항, 지배구조 10문항의 총 30문항으로 간소화한 중소기업 자가진단 체계를 별도로 운영하고 있다. 각 문항은 5점 리커트 척도(1점: 전혀 해당 없음 ~ 5점: 완전히 해당)로 측정되며, 본 연구의 독립변수가 바로 이 30문항 응답이다.

간소화된 자가진단 체계는 접근성 측면에서 진일보한 형태이지만, 세 가지 구조적 한계를 공유한다. 첫째, 진단 결과의 등급화 자동화가 부재하다. 기업이 30문항에 응답하여 총점을 얻더라도 그 점수가 MSCI·한국ESG기준원 등 정식 평가 기준에서 어느 등급에 해당하는지를 자동으로 알 수 없다. 둘째, 업종별 비교 위치와 구체적 개선 방향이 제공되지 않는다. 중소기업이 동일 업종 내에서 상대적으로 어느 수준에 있는지, 어느 영역을 우선적으로 개선해야 하는지에 대한 맥락적 해석은 여전히 전문가의 개입을 요구한다. 셋째, 개선과제의 실행 방안이 제공되지 않는다. 무엇이 부족한가를 파악한 이

후 어떻게 개선할 것인가에 대한 구체적 로드맵은 별도의 전문 컨설팅을 통해서만 얻을 수 있다.

자가진단 체계의 또 다른 특징은 자기보고(self-report) 기반 측정이라는 점이다. MSCI·한국ESG기준원 등 정식 평가는 공시 보고서·감사 자료·뉴스 텍스트 등 외부 관찰 데이터를 사용하는 반면, 자가진단은 기업 내부 담당자가 직접 응답하는 주관적 평가이다. 이는 응답 편향(self-enhancement bias)이나 반응 왜곡(response distortion)의 위험을 수반하지만, 동시에 기업 내부의 실제 이행 수준을 가장 직접적으로 반영하는 데이터이기도 하다. 본 연구는 이 자기보고 데이터의 예측 타당성을 처음으로 실증하며, 외부에서 독립적으로 산정된 ESG 등급을 종속변수(ground truth)로 삼아 자가진단 응답과 정식 평가 등급 간의 통계적 관계를 규명한다.

[표 2-1] ESG 평가기관·자가진단 도구 유형별 비교

유형	평가기관	등급체계	평가방식	평가규모	중소기업 접근성
글로벌 정식평가	MSCI ESG Ratings	AAA~CCC 7등급	산업 상대평가; 공시·뉴스 등 외부 관찰 데이터	상장 대기업 중심	낮음(유료)
글로벌 정식평가	Sustainalytics (Morningstar)	5범주(Negligible~Severe, 낮을수록 우수)	미관리 리스크 정량화; 외부 관찰 데이터	1.5만 개사 이상	낮음(유료)
글로벌 정식평가	S&P Global ESG Score	0~100점(CSA)	산업별 설문+공개정보 모델링	1.3만 개사 이상	낮음(설문 부담)
국내 정식평가	한국ESG기준원 (KCGS)	S·A+·A·B+·B·C·D 7등급	증빙·공시 기반; 원점수 비공개	상장사 약1,090사	낮음(상장사 한정)
국내 정식평가	서스틴베스트	AA·A·BB·B·C·D·E 7등급	상대평가; 공개정보 기반(연2회)	약1,300사	낮음(상장·대형 비상장 중심)
자가진단형	K-ESG 가이드라인	61문항·100점 (자율 진단)	기업 자기보고(self-report)	무제한 (정부 무료 배포)	높음(무료·자율)
자가진단형	본 연구 시스템	30문항 → AI 예측(Low/Mid/High)	자기보고+ AI 등급화·맞춤 컨설팅	무제한 (온라인, 단가 미정)	높음(자동 등급화·개선안 제공)

주: 등급 체계·평가대상 규모는 2024~2025년 각 기관 공개자료 기준이며, 평가대상 규모는 평가 연도·회차에 따라 변동될 수 있다.

앞서 검토한 정식 ESG 평가 체계와 자가진단 체계의 차이는 [표 2-1]과 같이 유형별로 정리된다. 글로벌·국내 정식 평가기관은 외부 관찰 데이터에 기반한 신뢰도 높은 등급을 제공하지만, 평가대상이 상장 대기업에 집중되어 있고 비용·참여 부담으로 인해 비상장 중소기업 대다수는 정식 평가의 범위 밖에 놓인다. 반면 자가진단형 도구는 무료·자율 접근이 가능하여 중소기업 접근성이 높으나, 응답 결과를 정식 등급 수준으로 환산하거나 구체적 개선 방향으로 연결하는 기능은 제공하지 못한다. 본 연구 시스템은 자가진단의 높은 접근성을 유지하면서 AI 기반 등급 예측과 맞춤형 컨설팅을 결합함으로써, 두 체계 사이의 공백을 메우는 중소기업 ESG 브릿지 도구로 설계되었다.

제 3 절 머신러닝 기반 ESG 예측 연구

ESG 등급 예측에 머신러닝을 적용한 선행연구는 2020년 이후 급격히 증가하였으며, 주로 상장 대기업의 재무·공시 데이터를 활용한 회귀 및 분류 모델로 전개되어 왔다. Jiang(2024)[9]은 선형회귀, 랜덤포레스트, GBM을 비교하여 GBM이 ESG 점수 예측에 최고 성능을 보임을 확인하였으며, SHAP 분석으로 산업분류가 ESG 점수의 1순위 결정 요인임을 규명하였다.

Zhang & Zhao(2025)[10]는 XGBoost와 SHAP을 결합하여 중국 A주 상장기업(2013~2022)의 ESG 등급을 예측하고, 재무 성과, 환경, 사회, 지배구조의 4개 차원에 걸친 15개 지표로 Accuracy 0.910의 성능을 보고하였다. 이 연구는 어느 변수가 예측에 기여하는가라는 질문에 SHAP 값으로 답함으로써, ESG 예측 연구에 설명가능성(explainability) 차원을 도입하였다. 한편 Cai et al.(2025)[11]은 중국 건설업 기업을 대상으로 XGBoost+SHAP에 LLM과 지식그래프(KG)를 결합한 통합 평가 모델을 제안하고, SHAP 분석으로 환경(38.7%) 〉 거버넌스(32.0%) 〉 사회(29.3%) 순의 기여도 패턴을 보고하였다.

9) Jiang, X. (2024). Predicting corporate ESG scores using machine learning: A comparative study. Proceedings of the 3rd International Conference on Financial Technology and Business Analysis. https://doi.org/10.54254/2754-1169/118/2024.18728

10) Zhang, J., & Zhao, Z. (2025). Corporate ESG rating prediction based on XGBoost-SHAP interpretable machine learning model. Expert Systems with Applications, 295, Article 128809.

11) Cai, B., Ye, Z., & Chen, S. (2025). Intelligent ESG evaluation for construction enterprises in China: An LLM-based model. Buildings, 15(15), Article 2710.

[표 2-2] ML 기반 ESG 등급 예측 선행연구 비교

연구	데이터	기법	주요성능	한계
Jiang et al. (2024)	S&P500 ESG 점수, 재무지표	LR,RF·GBM + SHAP	GBM(R^2 0.3863, 회귀)	회귀 기반(분류 아님), 대기업 재무 데이터 한정
Cai et al. (2025)	중국 건설업 기업	XGBoost + SHAP	E>G>S 패턴	단일 산업(건설업) 한정
Zhang & Zhao (2025)	중국 A주 상장기업	XGBoost + SHAP	Acc 0.91, 4차원 15지표	재무지표 의존, 비상장 일반화 미검증
김현수 등 (2022)	국내 중소기업	자가진단 체계 제안	-	진단 제시에 그침
본 연구	비상장 중소기업 299개사 자가진단	GBM + SHAP + 파생신호	Acc 0.8631, F1 0.8544	자기보고 데이터 예측력 실증

제 4 절 LLM과 ESG 컨설팅 자동화

대규모 언어모델(Large Language Model, LLM)을 ESG 영역에 적용한 연구는 2022년 이후 급증하였다. 이는 OpenAI의 ChatGPT, GPT-4 공개와 Anthropic의 Claude 시리즈 등장으로 LLM의 도메인 특화 활용 가능성이 실무 수준에서 입증된 이후의 흐름이다. ESG 맥락에서 LLM의 강점은 비정형 텍스트(공시 보고서, 법규, 뉴스, 가이드라인)에서 구조화된 의미를 추출하고, 맥락에 적합한 개선 방안을 자연어로 생성할 수 있다는 점이다. 다만 LLM은 환각(hallucination), 일관성 결여, 도메인 특수 지식 부족 등의 한계를 가지며, 이를 보완하기 위한 RAG와 KG 기반 접근이 활발히 연구되어 왔다.

He et al.[12]의 ESGenius는 EMNLP 2025에서 발표된 대표적 연구로, 50개 LLM의 ESG 지식을 체계적으로 벤치마킹하였다. 특히 Zero-shot 설정 대비 RAG 방식이 ESG 질의응답의 정답률을 평균 15% 이상 향상시킴을 실

12) He, C., Zhou, X., Wu, Y., et al. (2025). ESGenius: Benchmarking LLMs on Environmental, Social, and Governance (ESG) and Sustainability Knowledge. In Proceedings of EMNLP 2025 (pp. 14612–14653). arXiv:2506.01646

증하였으며, 이는 외부 지식 검색이 LLM의 도메인 지식 부족을 효과적으로 보완함을 보여준다. Wu et al.[13]의 SusGen-GPT는 NAACL 2025에서 발표된 연구로, ESG 보고서 자동 생성에 특화된 LLM을 개발하고 구체성, 실행 가능성 지표의 유용성을 실증하였다. 이 연구는 LLM이 범용 모델의 지식만으로는 전문 ESG 보고서를 생성하기 어렵다는 점과, 도메인 특화 파인튜닝 또는 맥락 주입이 필수적임을 뒷받침한다.

RAG는 Lewis et al.[14]이 제안한 프레임워크로, 외부 지식을 검색하여 LLM의 입력 맥락에 포함시킴으로써 환각을 감소시키고 최신 정보를 반영한다. 이 기본 아키텍처는 지식 베이스 구축, 벡터 임베딩과 유사도 검색, 검색된 문서를 맥락에 주입한 LLM 호출로 이루어지며, 오늘날 대부분의 도메인 특화 LLM 애플리케이션의 표준으로 자리 잡았다. Karakurt와 Akbulut[15]의 체계적 문헌검토에 따르면, 63개 RAG 연구 중 80.5%가 FAISS(Facebook AI Similarity Search) 등 밀집 벡터 검색(dense vector retrieval) 라이브러리를 표준 검색 방식으로 채택하였으며, 이는 FAISS가 대규모 벡터 데이터에 대한 근사 최근접 이웃(ANN) 검색에서 우수한 성능과 오픈소스 가용성을 제공하기 때문이다.

Edge et al.[16]의 GraphRAG는 KG를 RAG에 결합하여 전역적 쿼리 처리(global query)와 연쇄 추론(multi-hop reasoning)을 가능하게 한 중요한 기여이다. 기존 RAG가 의미적으로 유사한 문서 청크(chunk)를 평면적으로 검색하는 반면, GraphRAG는 개체(entity)와 관계(relation)의 그래프 구조를 활용하여 다단계 추론 경로를 추출한다. 예컨대 제조업 기업의 환경 규제 관련 개선과제를 질의할 때, 평면 RAG는 환경 규제 관련 문서를 반환하는 반면 GraphRAG는 제조업 → 대기환경보전법 → 위반 시 과태료 → ISO

13) Wu, Q., Xiang, X., Huang, H., et al. (2025). SusGen-GPT: A Data-Centric LLM for Financial NLP and Sustainability Report Generation. In Findings of NAACL 2025. arXiv:2412.10906

14) Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

15) Karakurt, E., & Akbulut, A. (2026). Retrieval-augmented generation (RAG) and large language models (LLMs) for enterprise knowledge management and document automation: A systematic literature review. Applied Sciences, 16(1), 368.

16) Edge, D., Trinh, H., Cheng, N., et al. (2024). From local to global: A graph RAG approach to query-focused summarization. arXiv:2404.16130

14001 인증 권고 → 관련 증빙 서식의 연쇄 경로를 제공한다. Liu[17]의 KG-RAG 연구는 이러한 듀얼 채널(vector + graph) 검색이 환각 억제와 답변 일관성 향상에 기여함을 정량적으로 실증하였다.

본 연구는 위 선행연구의 성과를 통합하여, RAG(FAISS 벡터 검색) + KG(Neo4j 지식그래프)의 듀얼 채널 맥락 강화를 중소기업 ESG 컨설팅에 적용한다. 본 연구의 방법론적 기여는 (1) 기존 RAG 연구가 대기업 공시 문서를 주 대상으로 한 반면, 본 연구는 중소기업 30문항 자가진단 응답과 K-ESG 가이드라인, 업종별 법규, ISO 인증 기준을 연결하는 도메인 특화 듀얼 채널을 구축하였다는 점, (2) 3개 맥락 조건(Baseline, RAG, RAG+KG) × 2개 LLM 모델(Haiku, Sonnet)의 3×2 완전교차 설계로 맥락 강화 효과를 체계적으로 검증한다는 점에 있다.

[표 2-3] LLM 기반 ESG·RAG·KG 선행연구 요약

연구	발표처/연도	핵심 기여	한계 및 확장 지점
Lewis et al. (2020)	NeurIPS 2020	RAG 프레임워크 제안	범용QA, 도메인 특화 부족
Edge et al. (2024)	arXiv 2024	GraphRAG 전역 질의	대규모 문서 대상, 중소기업 미적용
KG-RAG (2025)	Nature Sci. Rep.	듀얼 채널 환각 억제	범용, ESG 특화 부재
Zou et al.[18] (2024)	arXiv 2023	LLM+RAG 기반 ESG 데이터 추출	대기업 공시 중심
He et al. (2025)	EMNLP 2025	50개 LLM ESG 벤치마크	질의응답에 한정, 컨설팅 생성 부재
Wu et al. (2025)	NAACL 2025	SusGen-GPT 보고서 생성	대기업 공시 보고서 대상
본 연구	-	RAG-KG 결합 × 모델 3×2 완전교차 중소기업 ESG 컨설팅	Agentic AI 통합 파이프라인

17) Liu, W. (2025). Research on the construction and application of retrieval enhanced generation (RAG) model based on knowledge graph. Scientific Reports, 15, Article 40425.

18) Zou, Y., Shi, M., Chen, Z., Deng, Z., Lei, Z., Zeng, Z., Yang, S., Tong, H., Xiao, L., & Zhou, W. (2023). ESGReveal: An LLM-based approach for extracting structured data from ESG reports. arXiv preprint arXiv:2312.17264.

제 5 절 Agentic AI의 개념과 구조 및 ESG 적용

1. Agentic AI의 정의와 핵심 구성 요소

Agentic AI는 자율적으로 환경을 인식하고, 추론하며, 계획을 수립하고, 도구를 활용하여 실행하는 AI 시스템을 의미한다. 기존의 단일 프롬프트-응답 방식(single-turn)과 달리, Agentic AI는 인식(Perceive), 추론(Reason), 계획(Plan), 실행(Act)의 반복적 사이클을 통해 복잡한 과업을 단계적으로 수행한다. Wang et al.의 선행 연구는 LLM 기반 자율 Agent의 4대 핵심 요소로 Memory(기억), Planning(계획), Tool-Use(도구 활용), Action(실행)을 제시하였으며, 이 요소들이 상호작용하는 방식이 Agent의 능력을 결정한다고 논증하였다. Xi et al.[19]은 arXiv에 공개된 포괄적 개관 연구에서, LLM 기반 Agent의 잠재력을 단일 Agent(single-agent), 멀티 Agent(multi-agent), 인간과 Agent의 협업(human agent collaboration)의 세 유형으로 분류하였다.

Agentic AI의 위치를 명확히 하기 위해서는 LLM, RAG, AI Agent, Agentic AI를 단계적으로 구분할 필요가 있다. 단일 LLM은 학습된 파라미터에 내재된 지식만으로 입력 프롬프트에 대한 응답을 생성하며, 외부 도구나 지속적 기억을 갖지 않는다. RAG는 응답 생성에 앞서 외부 문서를 검색하여 맥락을 보강함으로써 사실성을 높이지만, 검색과 생성에 머무를 뿐 외부 환경에 대한 행위(action)는 수행하지 않는다. AI Agent는 여기에 도구 사용 능력과 자율적 의사결정을 더하여, 주어진 목표를 인식, 계획, 실행, 관찰의 순환으로 분해하고 도구 호출 결과를 관찰하며 과업이 완료될 때까지 반복한다. Agentic AI는 이러한 단일 Agent를 넘어, 서로 다른 역할을 가진 복수의 전문화된 Agent가 상위 조율자(orchestrator)의 통제 아래 협업하여 하나의 복합 과업을 분담 수행하는 시스템을 가리킨다. 즉 LLM에서 Agentic AI로의 이행은 생성에서 검색 보강으로, 다시 자율적 행위로, 최종적으로 다중 Agent 협업으로 능력이 누적되는 과정으로 이해할 수 있다.

실무적으로 이 발전 경로는 여섯 단계로 정리된다. 첫째는 프롬프트에만

19) Xi, Z., Chen, W., Guo, X., et al. (2023). The rise and potential of large language model based agents: A survey. arXiv:2309.07864

응답하는 초기 LLM, 둘째는 긴 맥락을 처리하는 롱컨텍스트 LLM, 셋째는 웹 검색·코드 실행 등 외부 도구를 호출하는 도구 사용 LLM, 넷째는 텍스트를 넘어 이미지·음성을 처리하고 정보를 기억하는 멀티모달·기억 단계, 다섯째는 목표를 받아 계획·도구 선택·실행·점검을 자율적으로 수행하는 추론 Agent, 여섯째는 오케스트레이터가 다수의 전문 Agent와 기억·도구·승인 절차를 결합하여 업무 전체를 실행하는 Agent 팀 시스템이다. 본 연구의 시스템은 예측·컨설팅·검증의 역할을 분담하는 복수의 Agent가 협업한다는 점에서 다섯째 단계에서 여섯째 단계로 이행하는 구간에 위치한다.

Agentic AI 시스템을 구성하는 핵심 개념은 다음과 같이 정리된다. 상위 조율자인 오케스트레이터는 목표를 하위 과업으로 분해하여 특정 과업만 수행하는 서브 Agent(subagent)에 위임하며, 각 Agent는 인식·계획·실행·관찰로 이어지는 Agent 루프(agent loop)를 따른다. Agent가 외부 함수·API·코드 실행기를 호출하여 실제 환경에 작용하는 능력을 도구 사용(tool use)이라 하고, 도구와 데이터 원천을 표준 인터페이스로 연결하는 개방형 규약을 모델 컨텍스트 프로토콜(MCP)이라 한다. 메모리는 단일 상호작용 내에서 유지되는 단기 기억과 외부 저장소에 보존되는 장기 기억으로 구분되며, 한 번에 읽고 추론할 수 있는 입력의 한계를 컨텍스트 윈도(context window)라 한다. 출력의 신뢰성과 안전성을 확보하기 위한 장치로는 출력을 검증된 외부 데이터에 결부시켜 환각을 줄이는 그라운딩(grounding), 허용되지 않는 행위를 차단하는 가드레일(guardrails), 호스트 시스템에 영향을 주지 않고 코드를 실행하는 샌드박싱(sandboxing), 그리고 위험도가 높은 행위 이전에 인간의 확인을 요구하는 인간 개입(Human in the Loop) 게이트가 있다. 이처럼 다수의 전문 Agent가 협력하여 단일 Agent보다 빠르고 신뢰성 있게 복합 목표를 달성하는 구조를 멀티 Agent(multi-agent) 시스템이라 한다.

이러한 Agent의 행동은 통상 여러 계층의 시스템 프롬프트로 규정된다. Agent의 정체성, 핵심 임무, 대화 규칙, 사용 가능한 도구 규칙, 기억·데이터 활용 규칙, 행동의 경계, 위임·이관 규칙, 안전·준법 규칙의 계층이 그것이며, 실제 대화에서 드러난 문제를 반영하여 프롬프트를 반복적으로 개선하는 순환 구조가 Agent의 신뢰성을 높인다. 다중 Agent의 조직 형태는 크게 두 가

지로 구분된다. 하나는 동등한 Agent들이 상호 정보를 주고받으며 협의하는 Agent 팀(agent teams) 방식이고, 다른 하나는 오케스트레이터가 다수의 과업을 병렬로 분기시킨 뒤 각 과업에서 구현, 검증, 수정 역할의 Agent가 순차적으로 작동하고 그 결과를 다시 취합하는 동적 워크플로(dynamic workflows) 방식이다. 본 연구의 파이프라인은 예측 결과가 컨설팅의 입력으로, 컨설팅 산출물이 다시 검증의 대상으로 전달되는 순차적 위임 구조를 채택하여 후자에 가깝다.

Hong et al.[20]의 MetaGPT는 ICLR 2024에서 발표된 SOP(Standard Operating Procedure) 기반 멀티 Agent 협업 프레임워크로, 소프트웨어 개발 프로세스에서 PM, 아키텍트, 엔지니어 역할을 각기 다른 LLM Agent에 할당하고 각 역할의 산출물(PRD, 기술 설계서, 코드)을 구조화된 흐름으로 전달하는 방식을 제안하였다. 이 연구는 각 Agent가 동일한 LLM을 사용하더라도 역할 분담과 산출물 체계화만으로 단일 LLM 호출보다 훨씬 우수한 결과를 얻을 수 있다는 점을 실증하였으며, 본 연구의 ESG 등급 예측 Agent, LLM 컨설팅 Agent 분할 설계의 이론적 근거를 제공한다. Khamis et al.[21]은 IEEE Access에 Agentic AI 시스템의 아키텍처와 평가 체계를 제시하여, 다수의 전문 Agent로 구성된 시스템에서 Agent 간 상호작용과 협력 패턴을 다차원으로 평가하는 틀을 제안하였다. Gartner[22]는 2026년 말까지 기업용 애플리케이션의 40%가 특정 업무에 특화된 AI Agent를 통합할 것으로 전망하였으며, 이는 2025년 기준 5% 미만에서 크게 증가한 수치로서 Agentic AI가 학술적 논의를 넘어 실무 배포 단계로 이행하고 있음을 시사한다.

이러한 다중 Agent 구조를 설계하는 방식은 정형화된 패턴으로 정리할 수 있다. Anthropic[23]은 효과적인 Agent 구축에 관한 기술 문서에서, 검색·도구·기억으로 능력을 보강한 증강 LLM(augmented LLM)을 기본 구성 요소로

20) Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Zhang, C., et al. (2024). MetaGPT: Meta programming for a multi-agent collaborative framework. In Proceedings of ICLR 2024 (pp. 1–20).

21) Khamis, A., Karray, F., & Kamel, M. (2025). Agentic AI systems: Architecture and evaluation. IEEE Access, 13, 45780–45795.

22) Gartner (2025). Predicts 2026: Agentic AI will reshape enterprise software. Gartner Research.

23) Anthropic (2024a). Building effective agents. Anthropic Engineering.

삼고, 그 위에서 작동하는 시스템을 사전에 정의된 경로를 따르는 워크플로(Workflow)와 모델이 스스로 과정을 주도하는 자율 Agent(agent)로 구분하였다. 워크플로는 다시 다섯 가지로 나뉜다. 하나의 과업을 순차 단계로 분해하여 앞 단계의 출력을 다음 단계의 입력으로 넘기는 프롬프트 체이닝(Prompt Chaining), 입력을 유형별로 분류하여 적합한 처리 경로로 보내는 라우팅(Routing), 서로 독립적인 하위 과업을 동시에 처리하는 병렬화(Parallelization), 상위 조율자가 하위 과업을 동적으로 분해하여 여러 작업자 Agent에 위임하고 결과를 종합하는 오케스트레이터 워커(orchestrator workers), 생성과 평가를 분리하여 평가 결과를 바탕으로 출력을 반복 개선하는 평가자 최적화(evaluator optimizer)가 그것이다. 자율 Agent는 이러한 정형 경로 없이 목표만 부여받아 계획, 실행, 관찰, 조정을 스스로 반복한다. 단일 LLM 호출에 해당하는 단일 호출 방식까지 포함하면, Agentic AI의 설계 패턴은 [그림 2-1]과 같이 여덟 가지로 정리된다.



[그림 2-1] Agentic AI 워크플로 8대 패턴과 본 연구의 적용 범위

본 연구의 시스템은 이 가운데 네 가지 패턴을 결합한 형태이다. 증강 LLM(⑥)을 기본 단위로 삼아 RAG와 KG로 응답을 그라운딩하며, 자가진단 예측을 담당하는 ESG 등급 예측 Agent와 개선 컨설팅을 담당하는 LLM 기반 ESG 컨설팅 Agent를 거쳐 검증 단계로 이어지는 순차 위임은 프롬프트 체이닝(②)에 해당한다. 상위 오케스트레이터가 예측·컨설팅·검증의 하위 과업

을 분해하여 각 Agent에 위임하고 산출물을 종합하는 구조는 오케스트레이터-워커(④) 패턴을 따른다. 생성된 컨설팅을 자동 채점과 전문가 검수의 이원 절차로 평가하여 품질 기준 미달 건을 선별 개선하는 Stage 4의 검증 체계는 평가자-최적화(⑤)에 인간 개입(HITL)을 결합한 형태이다. 반면 목표만 부여받아 과정 전체를 스스로 주도하는 완전 자율 Agent(⑧)는 Stage 4의 HITL 프로토타입 이후 단계에 해당하며, 본 연구에서는 후속 연구 과제로 위치한다. 단일 호출(①)은 비교 기준선으로만 활용되고, 라우팅(③)과 병렬화(⑦)는 본 연구의 설계 범위에 포함되지 않는다.

이상의 네 패턴이 본 연구의 파이프라인에서 어떻게 결합되어 자가진단 입력부터 보고서·증빙 산출까지 이어지는지를 단계별로 도식화하면 [그림 2-2]와 같다.



[그림 2-2] 본 연구 파이프라인의 Agentic AI 패턴 매핑

2. HITL(Human in the Loop) 선행연구

Human in the Loop(HITL)은 AI 시스템의 의사결정 파이프라인에 인간 전문가의 판단을 통합하는 설계 원칙이다. IBM Think[24]는 HITL을 자동화 시스템의 운영·감독·의사결정 과정에 인간이 참여하여 정확성·안전성·책임성·윤리적 판단을 보장하는 방식으로 정의하고, 그 구현 방식으로 지도학습, 인간 피드백 기반 강화학습(RLHF), 능동학습(active learning)을 제시하였다. ESG 컨설팅과 같은 고위험 의사결정 영역에서는 법규 정확성·실행 적합성 등

24) IBM Think (2025). What is Human in the Loop? IBM Think.

의 보장 때문에 HITL 방식이 필수적이다.

Parseur[25]는 고위험 문서·의사결정 업무에서 AI 자동화에 인간 검토를 결합한 HITL 방식이 AI 단독 운영보다 정확도와 신뢰성을 높인다고 정리하였다. Gartner에 따르면 2025년까지 신규 법률 기술 자동화 솔루션의 30%가 소프트웨어와 인력을 결합한 인간 개입(HITL) 방식으로 제공될 것으로 전망된다. SuperAnnotate는 LLM-as-Judge 단독 평가가 편향과 한계를 가지므로 인간 검토를 병행하는 것이 바람직함을 강조하였으며, 이는 본 연구가 AI 자동채점과 전문가 15인 평가를 병행하는 이원적 검증 체계를 채택한 직접적 근거이다. LangChain[26]은 실무에 배포되는 AI Agent에서 쓰기·삭제와 같은 중요한 작업에 대해 인간 승인을 요구하거나 Agent 권한을 읽기 전용으로 제한하는 통제 방식이 적용되고 있음을 보고하였으며, 이는 HITL이 단순한 이론적 원칙이 아니라 실무 배포의 현실적 요구임을 보여준다.

3. 본 연구와 선행연구의 차별성

본 연구는 위에서 검토한 네 영역(중소기업 ESG, LLM 기반 ESG, Agentic AI, HITL)의 교차점에서 체계적으로 구성되며, 이를 통해 기존 연구들이 개별적으로 제시한 기여를 통합적 프레임워크로 재구성한다. 선행연구 대비 본 연구의 차별성은 다음 여섯 차원에서 확인된다.

첫째, 데이터 측면에서 기존 연구는 주로 상장 대기업의 재무·공시 데이터(Cai 2025, Wang 2025)에 의존한 반면, 본 연구는 비상장 중소기업 299개사가 직접 응답한 30문항 자가진단 자기보고 데이터를 활용한다. 둘째, AI 구조 측면에서 기존 연구가 단일 LLM 호출 기반의 single-turn 방식(ESGenius 2025, SusGen-GPT 2025)에 머문 데 비해, 본 연구는 ESG 등급 예측 Agent, LLM 기반 ESG 컨설팅 Agent, 전문가 검증 Agent가 협업하는 Agentic AI 5단계 멀티에이전트 파이프라인을 구성한다. 셋째, 맥락 강화 방식에서 대부분의 선행연구가 벡터 검색 단독에 의존한 반면, 본 연구는 RAG

25) Parseur (2025). Human in the Loop AI (HITL): Complete guide to benefits, best practices and trends for 2026. Parseur Blog.

26) LangChain (2024). State of AI agents report 2024. LangChain Blog.

와 지식그래프 연쇄추론을 결합한 KG-RAG 듀얼 채널을 적용한다. 넷째, 산출물 측면에서 기존 연구가 텍스트 보고서나 점수만을 출력한 데 비해, 본 연구는 진단결과 요약, 맞춤형 컨설팅, 증빙서식의 3종 산출물을 자동 생성한다. 다섯째, 검증 측면에서 기존 연구가 정확도, F1, ROUGE 등 학술적 성능 지표에 국한된 반면, 본 연구는 전문가 HITL 평가에 기반한 품질 검증과 AI 재현성 교차검증을 함께 수행한다. 여섯째, 품질평가 측면에서 기존 연구가 RAGAS, G-Eval, ESGenius와 같은 평면적 지표 집합을 사용한 데 비해, 본 연구는 상위 → 3관점 → 5차원으로 이어지는 EIEC 3층 계층 통합 프레임워크를 적용한다.

[표 2-4] 선행연구 대비 본 연구의 차별성 매핑

차별성 포인트	기존 선행연구	본 연구
데이터	상장 대기업 재무·공시 데이터(Jiang 2024, Cai 2025, Zhang & Zhao 2025)	비상장 중소기업 299개사 자가진단 30문항 자기보고
AI 구조	단일 LLM 호출, single-turn (ESGenius 2025, SusGen-GPT 2025)	Agentic AI 5단계 멀티에이전트 파이프라인(ESG 등급 예측 Agent·LLM 기반 ESG 컨설팅 Agent·전문가 검증 Agent)
RAG 방식	벡터 검색 단독 (대부분 선행연구)	RAG + 지식그래프 연쇄추론 결합(KG-RAG 듀얼 채널)
산출물	텍스트 보고서 또는 점수만 출력	진단결과 + 컨설팅 + 증빙서식 3종 자동 생성
검증	학술적 성능 지표만 (정확도, F1, ROUGE)	전문가 14인 HITL 기반 품질 검증+ AI 재현성 CV 검증 추가
품질평가	평면적 지표 집합 (RAGAS, G-Eval, ESGenius)	EIEC 3층 계층 통합 프레임워크(상위→3관점→5차원)

제 3 장 Agentic AI 기반 ESG 컨설팅 프레임워크

제 1 절 ESG 컨설팅 프레임워크 설계

1. Agentic AI 5단계 프레임워크

연구의 Agentic AI 프레임워크는 중소기업 ESG 컨설팅의 전체 프로세스를 5개 단계(Stage)로 구조화한다. 이 프레임워크는 Acharya et al.의 인식(Perceive), 추론(Reason), 계획(Plan), 실행(Act), 학습(Learn) 사이클과 Hong et al.의 SOP 기반 멀티 Agent 협업 구조를 ESG 컨설팅 도메인에 특화하여 재설계한 것이다. 5단계는 단순한 순차적 처리가 아니라, 각 단계가 다음 단계의 입력 맥락을 구조적으로 제공하는 페루프(closed-loop)를 구성하며, 필요한 경우 이전 단계로 피드백을 전달하여 전체 파이프라인의 품질을 반복적으로 개선한다.

[표 3-1] 5단계 ESG 컨설팅 프레임워크

Stage	단계명	핵심 기능	주요 기술	Agent/역할
1	자가진단 입력 (Perceive)	30문항 웹 입력, 기업 기본정보 수집, 데이터 유효성 검증	웹 프론트엔드, 입력 검증 로직	Orchestrator
2	ML 등급 예측(Reason)	GBM 3 클래스 분류, 파생 정량신호 산출, SHAP 해석	GBM, SHAP, Stratified 5-Fold CV	ESG 등급 예측 Agent
3	LLM 컨설팅 생성 (Plan+Act)	RAG-KG 결합 맥락 구성, 맞춤형 보고서 3종 자동 생성	RAG(FAISS), KG(Neo4j), Claude API	LLM 컨설팅 Agent
4	전문가 검증 (Verify)	HITL 전문가 검수, 법규 정확성 확인	HITL 인터페이스, 검수 워크플로우	전문가 검증 Agent (HITL)
5	모니터링·학습 (Learn)	ESG 등급 변화 추적, 모델 재학습	종단 데이터, MLOps 파이프라인	Monitoring Agent

주: Stage 1~3이 본 연구의 구현·검증 범위이며, Stage 4는 전문가 15인 평가를 통한 개념 실증, Stage 5는 향후 연구 방향으로 프레임워크 수준의 설계를 제시한다.

Stage 1은 파이프라인의 시작인 자가진단 입력 단계로, 기업 담당자가 30문항 자가진단 웹 인터페이스에 응답을 입력한다. 이 단계의 핵심 기능은 단순한 데이터 수집이 아니라 데이터 유효성 검증(range check, consistency check)과 기업 메타데이터(업종, 규모, 지역) 표준화이다. Orchestrator 역할은 후속 Agent들이 일관된 입력 포맷을 받을 수 있도록 데이터를 정제하는 것이다.

Stage 2는 ML 기반 등급 예측 단계로, ESG 등급 예측 Agent가 자가진단 응답을 독립변수로 받아 Gradient Boosting Classifier로 3클래스 ESG 등급을 예측한다. 예측 결과는 확률분포 형태로 출력되며, 가장 높은 확률을 가진 클래스가 예측 등급이 된다. 동시에 SHAP 분석을 통해 예측에 기여한 상위 피처와 그 기여 방향(강점/약점)을 해석한다.

Stage 3은 LLM 컨설팅 생성 단계로, LLM 컨설팅 Agent가 Stage 2의 예측 결과와 SHAP 해석을 입력 맥락으로 받아 맞춤형 ESG 컨설팅 보고서를 생성한다. 이 단계에서 RAG 채널은 벡터 검색을 통해 관련 법규, 인증, 모범 사례 문서를 검색하고, KG 채널은 Neo4j 그래프에서 취약 항목 → 법규 → 인증 → 개선과제 → 증빙의 연쇄 경로를 추출한다. 두 채널의 결과가 LLM 프롬프트에 통합되어 Claude Sonnet API를 호출하며, 출력은 진단결과 요약, 맞춤형 컨설팅, 증빙서식 가이드의 3종 보고서로 구조화된다.

Stage 4는 전문가 검증 단계로, LLM 컨설팅 Agent의 출력을 ESG 전문가가 검수하며, 본 연구에서는 전문가 15인 평가를 통해 개념적 타당성을 실증한다.

Stage 5는 모니터링·학습 단계로, 배포 이후 기업의 ESG 등급 변화와 AI 컨설팅의 효과를 종단적으로 추적하여 모델을 재학습한다.

2. 연구 범위

본 연구는 5단계 프레임워크 중 Stage 1~3을 구현·검증 범위로 설정하고, Stage 4는 개념 실증 수준에서 수행하며, Stage 5는 프레임워크 수준의 설계 제안으로 다룬다. 이러한 범위 설정의 타당성은 다음 세 가지 근거에 기반한다.

첫째, Stage 1~3은 Agentic AI 파이프라인의 핵심 추론, 실행 루프(Perceive, Reason, Act)를 완성하는 최소 필수 범위이다. 이 세 단계가 있으면 자가진단 입력부터 맞춤형 컨설팅 산출까지의 완전한 엔드 투 엔드(end to end) 서비스가 구현되며, 이 자체로 독립적 검증과 배포가 가능하다. 둘째, ESG 등급 예측 Agent와 LLM 기반 ESG 컨설팅 Agent는 서로 다른 성격의 평가 지표로, 독립적으로 검증할 수 있다. ESG 등급 예측 Agent는 Accuracy, F1, Precision, Recall 등 분류 모델의 표준 지표로, LLM 기반 ESG 컨설팅 Agent는 EIEC 5차원 품질 지표로 각각 평가 가능하므로, 연구 설계의 명료성이 확보된다. 셋째, Stage 4·5의 본격적 구현은 각각 HITL 인터페이스 개발과 1~2년 이상의 종단적 데이터 축적을 요구하므로, 본 연구의 시간적 범위를 초과한다. 따라서 본 연구는 Agentic AI의 추상적 비전을 구체적 구현 사례로 착지시킨 기본 구현(baseline implementation)으로 위치하며, 이는 Stage 4·5의 고도화로 점진적으로 확장될 수 있는 체계적 기반을 제공한다.

제 2 절 LLM 기반 Agentic AI 구성 및 설계

1. 연구 내용 및 분석 방법

본 연구의 분석은 1장 2절에서 제시한 전체 프레임워크를 따른다. Agentic AI 5단계 프레임워크를 토대로 세 서브 Agent가 역할을 분담하며, 각 Agent의 검증이 곧 연구 내용에 대응한다. ESG 등급 예측 분석은 Stage 2의 ESG 등급 예측 Agent 성능을, 맥락 강화 품질 분석은 Stage 3의 LLM 기반 ESG 컨설팅 Agent 품질을, 품질 및 연산 효율성 분석은 두 Agent를 묶은 파이프라인 전반의 효율성을 검증하며, Stage 4의 전문가 검증 Agent는 6장 전문가 평가로 타당성을 개념 실증 수준에서 확인한다.

[그림 3-1]은 Agentic AI 5단계 프레임워크 구조도로, 해당 프레임워크에서 시스템 구현을 위하여 세 서브 Agent가 구현모듈(A1~A6)로 여섯가지 서브 Agent로 어떻게 역할을 분담을 하는지 나타내었다. 세 서브 Agent는 독립적으로 실행되는 것이 아니라, 상위 오케스트레이터가 각 Agent 내부의 구

현 모듈을 순서대로 호출하고 그 출력을 다음 단계의 입력으로 전달하며 피드백을 환류시키는 구조로 작동한다. 처리 흐름은 다음과 같다. 사용자가 자가진단에 응답하면 오케스트레이터는 이를 ESG 등급 예측 Agent에 전달하고, ESG 등급 예측 Agent는 예측 등급·E/S/G 영역별 기여도·취약 영역을 산출한다. 이 출력은 세 가지 경로로 LLM 기반 ESG 컨설팅 Agent의 입력 맥락으로 전환된다. 첫째, 취약 영역의 SHAP 설명은 어떤 항목의 모범사례와 법규를 검색할지를 결정하는 RAG 질의 맥락이 되고, 둘째, 예측 등급과 신뢰도는 컨설팅의 강도(낮은 등급일수록 기초 과제 우선)를 조정하며, 셋째, 업종 분류는 KG 탐색을 해당 업종으로 한정하는 필터로 작용한다.



[그림 3-1] Agentic AI 5단계 프레임워크 구조도

LLM 기반 ESG 컨설팅 Agent는 이 맥락을 받아 컨설팅 보고서를 생성하고, 그 산출물은 다시 전문가 검증 Agent의 검증 대상으로 전달되어 EIEC 5차원과 전문가 평가의 결과가 ESG 등급 예측 Agent와 LLM 기반 ESG 컨설팅 Agent의 개선 근거로 환류된다. 이로써 세 개의 Agent는 독립 모듈의 단순 결합이 아니라 하나의 유기적 파이프라인으로 연결된다.

위에서 기술한 세 개의 Agent와 여섯 구현 모듈의 관계는 개념층과 구현층의 2계층 구조로 정리된다. 즉 예측, 컨설팅, 검증의 세 개념 Agent가 상위 역할을 정의하고, 각 Agent 내부는 다시 세부 기능 단위인 구현 모듈로 나뉜

다. 구체적으로 ESG 등급 예측 Agent는 A1(진단)과 A2(벤치마크)를, LLM 기반 ESG 컨설팅 Agent는 A3(과제생성), A4(근거인용), A5(리포트)를, 전문가 검증 Agent는 A6(검증)을 내부 구현 모듈로 포함한다. 이 매핑은 A1에서 A6에 이르는 순차 처리 흐름(진단 → 벤치마크 → 과제생성 → 근거인용 → 리포트 → 검증)을 세 개념 Agent의 역할 분담에 정렬하였다.

각 Agent의 프레임워크는 역할에 따라 상이하다. ESG 등급 예측 Agent는 Gradient Boosting 3클래스 분류와 SHAP 설명을 결합한 머신러닝 예측 프레임워크, LLM 기반 ESG 컨설팅 Agent는 Claude(Haiku·Sonnet)에 RAG 검색과 Neo4j KG 추론을 결합한 듀얼 채널 프레임워크, 전문가 검증 Agent는 EIEC 5차원과 전문가(HITL) 평가를 결합한 검증 프레임워크이다. 이상의 역할 분담과 연결 구조에 따른 연구 내용과 분석 방법은 [표 3-2]와 같다.

[표 3-2] 연구 내용 및 분석 방법 체계

Stage	Agent	내용	분석 방법
자가진단 단독 예측	ESG 등급 예측 Agent	자가진단 30문항으로 ESG 등급(3클래스) 예측, 산업분류 /E/S/G 평균 추가 효과	GBM + Stratified 5-Fold CV, Accuracy, F1-weighted
산업군별 ESG 차이	ESG 등급 예측 Agent	산업군별 ESG 수준 차이	Welch's ANOVA + Games-Howell 사후검정
파생신호 추가예측	ESG 등급 예측 Agent	파생 정량신호 7개 추가의 성능 향상 효과	GBM + 5-Fold CV, 조건 A vs C-1 비교
맥락강화 품질분석	LLM 컨설팅 Agent	RAG/KG 맥락 강화의 5차원 품질 개선	3×2 완전교차 설계, One-way ANOVA, Tukey HSD
품질/연산 효율성분석	LLM 컨설팅 Agent	품질·연산 효율성 관계	품질/1K token 효율성 지표, 조건×모델별 비교
전문가 검증	전문가 검증 Agent	산출물 품질 검증 및 EIEC 5차원 평가(n=15). HITL 개선 신호를 파이프라인에 환류	기술통계, 신뢰도 (Cronbach α·ICC), 5차원 비교
수요기업 만족도	— (수요자 설문)	구현된 서비스에 대한 주 사용 대상 (중소기업 ESG 담당자) 만족도·활용 의향 평가	기술통계 분석, 전문가 패널과 분리 집계

제 3 절 자가진단 문항 설계 및 중요도 분석

1. 자가진단 문항 설계

산업통상자원부의 K-ESG 가이드라인은 정보공시(P), 환경(E), 사회(S), 지배구조(G) 영역에 걸친 61개 진단항목으로 구성되어 있으며, 각 항목은 점검기준, 성과점검 방법, 배점, 추가설명을 포함하여 기업이 자율적으로 ESG 경영 수준을 진단하도록 설계되었다. 그러나 61개 진단항목을 모두 응답하고 해석하는 일은 전문 인력과 예산이 제한된 중소기업의 현실적 역량을 초과하는 부담이다. 이러한 한계를 보완하기 위해 본 연구는 K-ESG 가이드라인의 61개 진단항목 가운데 ESG 등급 변별에 핵심적인 항목만을 선별하여 간소화한 중소기업 특화 자가진단 체계를 채택하였다. 이는 진단 항목의 타당성을 공신력 있는 공공 지침이라는 기존 표준에 근거하게 하는 동시에, 중소기업의 응답 부담을 최소화하려는 설계 의도에 따른 것이다.

이에 본 연구는 국내 ESG 진단문항, 평가모형의 상위 표준인 산업통상자원부의 K-ESG 가이드라인 체계에 기반한 중소벤처기업진흥공단의 자가진단 체계를 토대로 삼되, K-ESG 가이드라인의 61개 진단항목을 모두 망라하는 대신 ESG 등급 변별에 핵심적인 항목만을 선별하여 간소화[27]하였다. 이는 항목이 많아 응답 부담이 큰 전체 진단 체계에서 핵심만을 추출함으로써, 진단의 타당성을 공신력 있는 공공 표준에 근거하게 하면서도 중소기업의 응답 부담을 최소화하려는 설계 의도에 따른 것이다.

대상 표본은 비상장 중소기업이다. 본 연구는 외부에서 독립적으로 ESG 등급이 산정된 중소기업을 표본으로 삼았는데, 이는 표본을 임의로 구성한 것이 아니라 이미 ESG 평가 결과가 존재하는 기업을 대상으로 했기 때문에 자가진단 응답과 외부 평가 등급 간의 관계를 검증할 수 있고, 그 등급을 종속변수(ground truth)로 활용함으로써 데이터의 신뢰성과 공신력을 확보할 수 있다는 근거에 기반한다. 정식 ESG 평가는 상장 대기업에 집중되어 있어 비상장 중소기업 대다수가 평가 범위 밖에 놓이는 현실을 고려할 때, 외부 평가

27) 중소벤처기업진흥공단(2022). 「중소기업ESG 자가진단 도구 개발 및 보급 사업 보고서」. 중소벤처기업진흥공단 정책연구실.

등급이 확보된 중소기업 표본은 그 자체로 희소하고 검증 가능한 자료적 가치를 지닌다.

이렇게 설계된 자가진단 문항은 프레임워크의 Stage 1(자가진단 입력, Perceive)에 대응한다. Stage 1은 단순한 데이터 수집이 아니라 응답의 유효성 검증과 기업 메타데이터(업종, 규모, 지역)의 표준화를 수행하여, 후속 Agent가 일관된 입력 포맷을 받도록 정제하는 관문 역할을 한다. 정제된 자가진단 응답은 Stage 2의 ESG 등급 예측 Agent로 전달되어 예측 등급과 취약 영역으로 변환되고, 이 결과가 다시 Stage 3의 LLM 기반 ESG 컨설팅 Agent의 입력 맥락이 됨으로써, 자가진단 문항은 곧 컨설팅 산출의 출발점으로 기능한다. 즉 자가진단 문항 설계는 진단 자체를 위한 도구일 뿐 아니라, 자가진단 → 등급 예측 → 맞춤형 컨설팅으로 이어지는 페루프 파이프라인의 첫 입력을 규정하는 설계 요소이다.

2. 자가진단 문항 중요도 분석

본 연구에서 정제된 자가진단 응답은 Stage 2의 ESG 등급 예측 Agent로 전달되어 예측 등급과 취약 영역으로 변환되고, 이 결과는 다시 Stage 3의 LLM 기반 ESG 컨설팅 Agent의 입력 맥락이 된다. 따라서 자가진단 문항은 곧 컨설팅 산출의 출발점으로 기능하며, 자가진단 문항 설계는 진단 자체를 위한 도구일 뿐 아니라 자가진단 → 등급 예측 → 맞춤형 컨설팅으로 이어지는 페루프 파이프라인의 첫 입력을 규정하는 설계 요소가 된다. 본 연구의 분석 방법은 이러한 페루프 파이프라인이 각 단계에서 의도대로 작동하는지를 확인하기 위해 다음과 같이 설계한다.

먼저 파이프라인의 첫 변환 단계인 ESG 등급 예측은 GBM을 주 분류기로 채택하고, 모델 평가는 Stratified 5-Fold 교차검증으로 수행한다. 이는 정제된 자가진단 응답만을 입력으로 하여 전문 평가기관이 부여하는 ESG 평가등급을 어느 정도까지 재현할 수 있는지를 검증함으로써, 자가진단 문항이 후속 컨설팅의 출발점으로 기능할 만큼 충분한 예측 정보를 담고 있는지를 확인하기 위한 것이다. 나아가 예측 결과의 설명가능성은 XGBoost 대리 모델

을 기반으로 한 SHAP 분석을 통해 확보하여, 어떤 자가진단 문항이 등급 예측과 취약 영역 도출에 실질적으로 기여하는지를 규명할 수 있도록 설계한다.

산업군별 ESG 등급의 집단 간 차이는 Welch's ANOVA로 검정한다. 본 연구의 5개 산업군은 제조업이 압도적으로 많고 서비스업과 도소매업이 각각 1개에 불과하여 표본 크기가 크게 불균등하게 구성되어 있으며, 이로 인해 등분산 가정을 충족하기 어렵다. Welch's F-검정은 등분산과 등표본 크기 가정을 요구하지 않으므로 이러한 표본 구조에 적합하다. 사후검정으로는 동일하게 두 가정을 요구하지 않는 Games-Howell 방법을 적용하며, 유의수준은 α = 0.05로 설정한다.

페루프의 마지막 단계인 LLM 컨설팅의 품질과 효율성에 대한 실험은 3개 맥락 조건(Baseline, RAG, RAG+KG)과 2개 LLM 모델(Claude Haiku 4.5, Claude Sonnet 4.6)을 교차한 3×2 완전교차 설계로 수행한다. 이 설계의 통계적 정당성은 세 가지로 정리된다. 첫째, 완전교차 설계는 맥락과 모델의 주 효과뿐 아니라 두 요인 간의 상호작용 효과까지 분리하여 추정할 수 있게 하며, 이를 통해 모델 규모가 클수록 맥락 강화의 효과가 더 큰가와 같은 상호작용 가설을 검증할 수 있다. 둘째, 동일한 30개 기업을 6개 조건 모두에 교차 적용함으로써 기업 수준의 개별차가 조건 간 비교에 미치는 영향을 통제한다. 셋째, 각 셀에 30건씩 배정하여 총 180건을 생성하므로 ANOVA의 검정력 요건(셀당 n≥15)을 충분히 만족하도록 설계한다. 조건 간 차이는 One-way ANOVA로, 사후검정은 Tukey HSD로 검정하며, 모델 간 차이는 Welch's t-test로 검정하고, 유의수준은 α=0.05로 설정한다.

마지막으로 연산 효율성은 품질 총점을 1,000 토큰으로 나눈 값으로 정의한다. 이 지표는 동일한 연산 자원, 즉 API 비용의 직접적 결정 요인인 토큰 투입 대비 얼마나 높은 품질의 컨설팅을 생성하는지를 측정하기 위한 것이다. 토큰 사용량은 Anthropic Claude API의 응답 메타데이터에서 추출하는 input tokens와 output tokens의 합으로 산정한다. 이를 통해 실무 배포 시 비용을 최소화하면서 일정 수준 이상의 품질을 보장하기 위해 어느 조건을 선택해야 하는지에 대한 운영 의사결정의 근거를 확보하고자 한다.

제 4 장 ESG 등급 예측 Agent

제 1 절 ESG 등급 예측 Agent 설계

1. ESG 등급 예측 Agent의 정의

ESG 등급 예측 Agent는 전체 Agentic AI 파이프라인의 출발점으로, 중소기업 자가진단 응답만으로 공인 평가기관 등급 수준을 근사하는 예측 기능을 담당하는 머신러닝 모듈이다. 본 Agent는 두 개의 하위 구현 모듈로 구성된다. 진단 모듈(A1)은 자가진단 응답을 입력받아 ESG 등급과 취약 영역을 산출하고, 벤치마크 모듈(A2)은 그 결과를 동종 업종, 등급 분포와 대비하여 상대적 위치를 부여한다. Agent의 동작은 네 단계로 이루어진다. 첫째, 자가진단 응답을 수집 및 검증한다. 둘째, Gradient Boosting 기반 3클래스 분류로 ESG 등급을 예측한다. 셋째, SHAP 분석으로 예측 기여 요인과 취약 영역을 해석한다. 넷째, 예측 등급, 확률분포, 취약 영역 정보를 LLM 기반 ESG 컨설팅 Agent의 입력 맥락으로 전달한다.



[그림 4-1] ESG 등급 예측 Agent의 구조와 동작 흐름

2. ESG 등급 예측 Agent의 분석모델

본 Agent의 주 분석모델로 Gradient Boosting Machine(GBM)을 채택하였다. 선정 근거는 세 가지이다. 첫째, GBM 계열은 선행 ESG 등급 예측 연구에서 일관되게 우수한 성능을 보였으며, 본 연구의 예비 비교에서도 트리 부스팅 계열이 우위를 보였다(구체적 수치는 아래 대안 모델 비교 항 참조). 둘째, 5점 리커트 순서형 응답과 산업분류 더미가 혼합된 입력 구조에서 트리 기반 앙상블은 피처 스케일에 민감하지 않고 범주형, 수치형 혼합에 강건하다. 셋째, 결측·이상치 내성이 높아 자가진단 데이터 특유의 응답 결측·극단값에 적합하다. 아울러 예측의 설명가능성을 확보하기 위해 SHAP(SHapley Additive exPlanations) 분석을 병행하여 개별 예측의 기여 요인과 취약 영역을 해석한다.

가. 예측 모델 및 하이퍼파라미터

예측 모델 및 하이퍼파라미터에 사용된 GBM은 약한 학습기인 의사결정나무를 순차적으로 결합하여 이전 모델의 예측 오류를 점진적으로 보정하는 부스팅 계열 앙상블 기법이다. 주 분류기 GBM의 하이퍼파라미터는 제한적 그리드 탐색(27개 조합)으로 선정하였으며, 최종 설정은 다음과 같다.

[표 4-1] GBM 하이퍼파라미터 탐색 범위 및 최종 설정

하이퍼파라미터	탐색범위	최종 설정
n_estimators	{100, 200, 300}	200
max_depth	{3, 4, 5}	4
learning_rate	{0.05, 0.1, 0.2}	0.1
subsample	–	0.8

주: random_state=42(재현성 고정 시드). 최종 설정은 상위 5개 조합과 ±0.005 이내로 특정 설정에 과의존하지 않음.

클래스 불균형(Class 2 = 5.0%)을 고려하여 Stratified K-Fold 교차검증(k=5)을 적용하였다. 각 폴드에서 학습(약 240개사)과 검증(약 60개사)을 수

행하고, 5회 반복의 평균과 표준편차로 성능을 보고한다. 층화 분할은 High 등급 15개사가 각 폴드에 3개씩 고르게 배분되도록 하여 소수 클래스에 대한 안정적 평가를 보장한다. 설명가능성 분석을 위해 XGBoost 대리 모델(surrogate model)을 별도로 학습하고 TreeSHAP으로 SHAP 값을 산출하였다. 이 방식은 GBM과 XGBoost가 매우 유사한 트리 부스팅 구조를 공유한다는 점에서 타당하며, 두 모델의 Feature Importance가 상위 순위에서 수렴함을 확인하여 해석의 타당성을 보강하였다.

나. 대안 모델 비교 및 하이퍼파라미터 선정 과정

대안 모델 비교 및 하이퍼파라미터 선정 과정은 주 분류기 선정에 앞서 조건 A(Baseline)·Stratified 5-Fold CV(기본 하이퍼파라미터)로 다섯 모델을 비교하였다.

[표 4-2] 대안 모델 비교 (조건 A, Stratified 5-Fold CV)

모델	Accuracy (조건 A, 5-Fold)
Logistic Regression	0.72 (기준선)
Random Forest	0.78
Gradient Boosting	0.78
XGBoost	0.81
LightGBM	0.81

Gradient Boosting을 최종 주 분류기로 선택한 근거는 scikit-learn 생태계와의 통합성(StratifiedKFold, GridSearchCV 등 재현성 및 유지보수)과 ESG 예측 선행연구와의 직접 비교 가능성이다. XGBoost는 SHAP 분석의 대리 모델로 병행하였다.

클래스 불균형 대응으로 SMOTE, 클래스 가중치, 언더샘플링을 예비 검토하였으나, SMOTE는 High 클래스(n=15) 합성 표본의 과적합(0.79), 언더샘플링은 정보 손실(0.75)을 보였고, 주 설계(Stratified K-Fold)가 가장 균형 잡힌 성능을 제공하여 이를 채택하였다.

3. ESG 등급 예측 Agent의 입력-처리-출력

이 30문항은 국내 ESG 진단문항의 상위 표준인 산업통상자원부 K-ESG 가이드라인에 기반한 중소벤처기업진흥공단의 자가진단 문항을 기반으로 하여, ESG 등급 변별에 핵심적인 항목을 본 연구가 선별 및 간소화한 것으로, 그 전체 구성은 다음과 같다.

[표 4-3] ESG 자가진단 30문항 및 대응 증빙 양식

번호	코드	평가문항	대응증빙방식
1	E-01	환경경영 방침이 서면으로 제정·승인되어 있습니까?	중소기업중앙회ESG 규정례_환경경영 지침(안)
2	E-02	환경목표와KPI가 연간 계획으로 수립되어 있습니까?	중소기업중앙회ESG 규정례_환경경영 지침(안)
3	E-03	최근12개월 환경 성과관리표를 보유하고 있습니까? (Scope1/2 등)	중소기업중앙회ESG 규정례_환경경영 지침(안)
4	E-04	대기배출시설을 보유합니까? 보유 시 자가측정·점검을 수행합니까?	중소기업중앙회ESG 규정례_대기오염물질관리 규정(안)
5	E-05	폐수배출 또는 위탁처리를 적정하게 관리합니까?	중소기업중앙회ESG 규정례_폐수배출관리 규정(안)
6	E-06	유해화학물질(MSDS) 관리체계를 보유하고 있습니까?	중소기업중앙회ESG 규정례_화학물질관리 규정(안)
7	E-07	폐기물 분류 및 처리(일반/지정, 위탁 포함)가 기준에 맞게 관리됩니까?	중소기업중앙회ESG 규정례_대기오염물질관리 규정(안)
8	E-08	연간 에너지 사용량 집계 및 절감활동을 수행합니까?	중소기업중앙회ESG 규정례_환경경영 지침(안)
9	E-09	Scope1/2 배출 인벤토리를 보유하고 있습니까?	온실가스환산Tool
10	E-10	환경 관련 연간 교육계획 및 이행을 수행합니까?	환경교육 일지 및 인명부 서식
11	S-01	인권경영 선언문/정책을 보유하고 공표했습니까?	중소기업중앙회ESG 규정례_인권경영 규정(안)
12	S-02	임직원 고충·신고·구제 절차가 운영되고 있습니까?	고충사항 접수처리대장 서식
13	S-03	안전보건방침 수립 및 위험성평가를 정기 수행합니까?	중소기업중앙회ESG 규정례_안전보건관리 지침(안)
14	S-04	비상대응 절차 및 모의훈련을 실시합니까?	제397호_비상대피 시설은 이렇게 하세요! 포스터
15	S-05	근로계약, 임금, 근로시간 관리가 법정 기준에 부합합니까?	표준근로계약서 및 친권자 동의서 서식
16	S-06	차별금지·평등정책을 수립·교육하고 있습니까?	중소기업중앙회ESG 규정례_인권경영 규정(안)
17	S-07	협력사 행동규범(CoC) 및 서약을 운영합니까?	협력사 행동규범(CoC) 및 준수 서약서(안)
18	S-08	연간 인권/안전/윤리 등 사회분야 교육을 수행합니까?	성희롱예방교육 일지 및 인명부 서식
19	S-09	지역사회 공헌 활동을 수행합니까?	중소기업중앙회ESG 규정례_사회공헌 지침(안)
20	S-10	개인정보 처리방침을 수립·공지하고 준수합니까?	중소기업중앙회ESG 규정례_개인정보보호 규정(안)
21	G-01	윤리강령 및 부패방지 정책을 수립·교육합니까?	중소기업중앙회ESG 규정례_윤리경영 지침(안)

22	G-02	ESG 책임자(또는 위원회)가 지정되어 정기 보고합니까?	-
23	G-03	법규 준수 및ESG 리스크 관리 절차를 운영합니까?	-
24	G-04	공정거래 준수 프로그램을 운영합니까?	중소기업중앙회ESG 규정례_공정거래 지침(안)
25	G-05	정보보안 정책 및 점검 절차를 운영합니까?	중소기업중앙회ESG 규정례_윤리경영 지침(안)
26	G-06	내부신고 제도 및 신고자 보호절차를 운영합니까?	중소기업중앙회ESG 규정례_윤리경영 지침(안)
27	G-07	개인정보보호 책임자 지정 및 접근·권한 관리를 수행합니까?	중소기업중앙회ESG 규정례_개인정보보호 규정(안)
28	G-08	ESG 관련 정책·절차·기록 문서의 버전관리 체계를 운영합니까?	-
29	G-09	ESG 정책/성과를 홈페이지 등 외부에 공시합니까?	-
30	G-10	임직원 대상 윤리/정보보안/개인정보 교육을 실시합니까?	중소기업중앙회ESG 규정례_개인정보보호 규정(안)

입력은 이 자가진단 응답(E/S/G 각 10문항, 5점 리커트)이며, 처리는 Gradient Boosting 3클래스 분류 모델을 Stratified 5-Fold로 학습하고 피처 조건(A(Baseline), B(+산업분류), C(+E/S/G 평균), C-1(+파생 정량신호))별 성능을 검증하는 단계로 구성된다. 출력은 예측 등급, 클래스별 확률분포, SHAP 기반 취약 영역 설명의 다차원 정보이며, 이는 LLM 기반 ESG 컨설팅 Agent의 입력 맥락으로 전달되어 두 Agent를 하나의 유기적 파이프라인으로 연결한다.

가. 데이터 수집 및 전처리

분석에 사용된 데이터는 비상장 중소기업 299개사의 자가진단 응답이다. 각 기업은 ESG 자가진단 30문항(환경 E 10문항, 사회 S 10문항, 지배구조 G 10문항)에 5점 리커트 척도로 응답하였으며, 종속변수인 ESG 등급은 외부에서 독립적으로 산정된 결과이다. 독립변수인 자가진단 응답과 종속변수인 외부 평가 등급은 방법론적으로 서로 독립된 원천에서 수집된 것으로, 이 방법론적 독립성이 본 연구 ESG 등급 예측 분석 검증의 신뢰도를 담보하는 핵심 조건이다.

데이터 전처리는 세 단계로 수행되었다. 첫째, 결측치 처리 단계에서는 독립변수 30문항의 결측은 해당 문항 평균값으로 대체하였다. 평균 대체는 리커트 척도 데이터에 일반적으로 적용되는 방법으로, 본 연구 데이터에서는

문항별 결측률이 3% 이내로 낮아 대체 방식이 결과에 미치는 영향은 제한적이다. 종속변수 결측 1건은 예측 과제의 특성상 학습 및 검증 모두에 활용 불가하므로 분석에서 제외하였다. 둘째, 종속변수 변환 단계에서는 외부 평가의 원래 7단계 등급(C, D, B-, B0, B+, A0, A+)을 Low(C, D), Mid(B0, B+), High(A0, A+)의 3클래스로 재분류하였다. 이 재분류의 근거는 (1) 원래 7단계를 그대로 사용하면 각 클래스의 관측치 수가 극도로 적어 학습이 불안정해지는 점, (2) 실무 컨설팅에서 등급을 하위·중위·상위로 인식하는 3분할이 가장 일반적이라는 점에서이다. 셋째, 산업분류 코드 처리 단계에서는 KSIC 5자리 코드를 5개 산업군(제조업, 건설업, ICT, 서비스, 기타, 도소매)으로 재범주화하고, 조건 B 이후 피처 세트에서 원-핫 인코딩(One-Hot Encoding)으로 변환하였다. 원-핫 인코딩은 K개의 범주를 가진 명목형 변수를 K개의 이진 변수(0/1)로 분리하여 표현하는 변환 기법으로, 산업군 변수를 정수 라벨(예: 제조업=1, 건설업=2, ICT=3)로 코딩할 경우 모델이 범주 간 순서 또는 거리 관계를 학습할 위험을 제거한다. 산업군은 본질적으로 명목 척도(nominal scale) 변수이므로 범주 간에 어떠한 순서나 크기 관계도 존재하지 않으며, 따라서 각 산업군이 독립된 차원으로 표현되어야 모델이 산업군별 ESG 패턴을 편향 없이 학습할 수 있다. 본 연구에서는 5개 산업군 각각에 대해 이진 변수가 생성되어 피처 세트에 포함되었다.

[표 4-4] 종속변수(ESG 등급) 3 클래스 분포

클래스	ESG 등급범위	기업 수	비율(%)	실무적 의의
0 (Low)	C, D 등급	88	29.4%	하위 등급군— 기본 요건 미충족
1 (Mid)	B0, B+ 등급	196	65.6%	중위 등급군— 보통 수준 이행
2 (High)	A0, A+ 등급	15	5.0%	상위 등급군— 체계적 이행
합계	-	299	100.0%	-

[표 4-4]에서 확인되는 바와 같이, Mid 등급이 전체의 65.6%를 차지하는 불균형 분포는 중소기업 ESG 수준이 중간 범위에 집중되는 현실을 반영한다. 이러한 클래스 불균형은 머신러닝 학습에서 다수 클래스에 편향된 예측을 유

도할 수 있는 구조적 위험 요인이다. 예컨대 모델이 모든 샘플을 Mid로 분류하기만 해도 65.6%의 Accuracy를 달성할 수 있다. 이러한 편향을 방지하기 위해 본 연구는 Stratified K-Fold 교차검증으로 각 폴드의 클래스 비율을 전체 데이터와 동일하게 유지하였고, Accuracy와 더불어 F1-weighted를 병행 보고하여 소수 클래스(High) 예측 성능이 전체 지표에 반영되도록 하였다.

나. 피처 엔지니어링 및 조건 설계

예측 성능에 대한 피처 유형의 기여도를 체계적으로 분석하기 위해, 4가지 조건(A, B, C, C-1)의 피처 세트를 단계적으로 확장하는 설계를 적용하였다. 이 단계적 확장은 어느 피처가 예측에 얼마만큼 기여하는가를 계층적으로 분리 평가할 수 있게 하는 방법론적 장치이다. 각 조건은 자가진단 원점수만의 조건 A(Baseline)에서 출발하여 산업분류(B), E/S/G 영역 평균(C), 파생 정량신호(C-1)를 더하는 방식으로 구성되며, 구체적 구성은 다음과 같다.

[표 4-5] 피처 조건 설계 (조건 A, B, C, C-1)

조건	추가피처	피처수
A (Baseline)	자가진단 30문항 원점수 (E·S·G 각 10)	30
B (+산업분류)	A + KSIC 5개 산업군 원-핫(4 더미)	34
C (+E/S/G 평균)	B + E/S/G 영역 평균 3개(e_mean, s_mean, g_mean)	37
C-1 (+파생 정량신호)	A + 파생 정량신호 7개(strong_count, weak_count, e_std, s_std, g_std, es_gap, sg_gap)	37

파생 정량신호 7개(strong_count, weak_count, e_std, s_std, g_std, es_gap, sg_gap)는 30문항 원점수를 2차 가공하여 산출한 지표이며, 본 연구에서 독자적으로 설계한 피처 엔지니어링 장치이다. strong_count는 30문항 중 4점 이상 문항의 개수로 기업의 ESG 강점 영역 총량을, weak_count는 2점 이하 문항의 개수로 취약 영역 총량을 측정한다. e_std, s_std, g_std는 각 E/S/G 영역 10문항의 표준편차로 영역 내 이행 일관성(internal consistency)을 포착한다. 편차가 크면 특정 문항만 선택적으로 이행하는 패턴을, 편차가

작으면 영역 전반에 걸친 균등한 이행 패턴을 의미한다. es_gap은 E와 S 영역 점수의 격차, sg_gap은 S와 G 영역 점수의 격차로, 영역 간 이행 불균형 정보를 포착한다. 이러한 파생 변수는 원점수가 포착하지 못하는 분포 패턴 정보를 모델에 제공함으로써 예측 성능을 향상시킬 수 있다는 가설에 근거한다.

제 2 절 ESG 등급 예측 Agent 실험결과

1. 조건별 예측 성능 비교

[표 4-6]의 결과는 단계적 피처 확장의 효과를 명확하게 보여준다. 조건 A(Baseline, 자가진단 30문항 원점수)의 정확도 0.8093은, 대기업 재무·공시 데이터 기반 선행 ESG 예측 연구와 달리 중소기업 자가진단 데이터를 활용하여 달성한 결과이다. 즉, 자가진단 30문항 원점수만으로 ESG 등급을 약 81% 정확도로 예측할 수 있음을 실증하며, 이는 자가진단이 재무 지표보다 ESG 실천 수준을 더 직접적으로 반영함을 시사한다. 대기업 대상 Cai et al.(2025)이 93.33%의 분류 정확도를 보고한 점을 고려하면, 본 연구에서도 정량 신호 추가를 통한 추가 성능 개선이 기대된다.

[표 4-6] 조건별 예측 성능 (GBM, Stratified 5-Fold CV, n=299)

조건 및 피처 구성	Accuracy	±std	F1-weighted	⊿Acc	±std	n
A:30문항만 (Baseline)	0.8093	0.0274	0.7962	0.0350	—	299/5
B:+산업분류(KSIC5 그룹)	0.8094	0.0326	0.7934	0.0375	+0.0001	299/5
C:+E/S/G 평균	0.8494	0.0323	0.8370	0.0363	+0.0401	299/5
C-1:+파생 정량신호 7개	0.8631	0.0281	0.8544	0.0299	+0.0537	299/5

조건 B(+산업분류)는 KSIC 5개 산업군을 원-핫 인코딩으로 추가하였으나 예측 성능이 거의 개선되지 않았다(+0.0001). 이는 일견 직관에 반하는 결과이나 체계적 해석이 가능하다. 30문항 응답 자체에 이미 산업 특성이 내재되어 있기 때문이다. 예컨대 제조업 기업은 환경 관련 문항(E-02 환경경영체계, E-07 폐기물관리)에서, ICT 기업은 개인정보보호 관련 문항(S-10)에서

각기 다른 응답 분포를 보인다. 따라서 명시적 산업분류 변수의 추가는 이미 문항 응답에 내재된 정보의 중복에 머무르며, 모델 학습에 새로운 신호를 거의 제공하지 못한다. 오히려 표준편차가 0.0274에서 0.0326으로 증가하여 예측 안정성이 다소 악화되었는데, 이는 추가된 범주형 더미가 일부 폴드에서 노이즈로 작용할 수 있음을 시사한다.

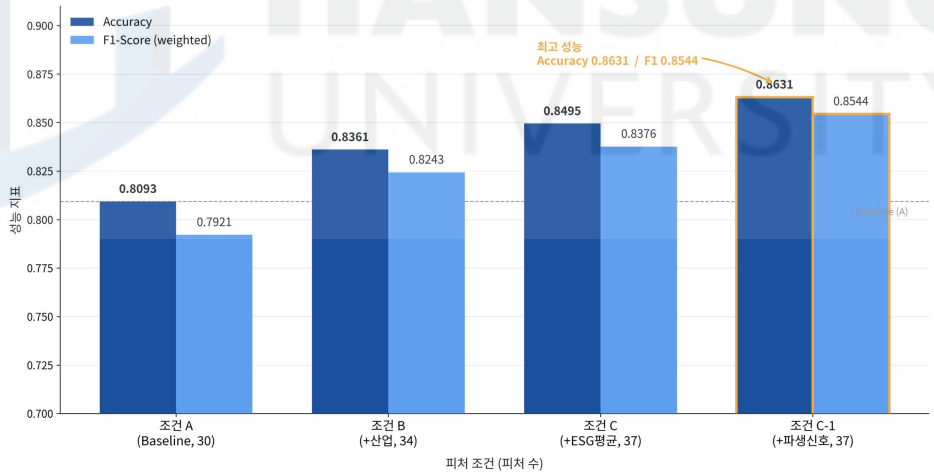
조건 C(+E/S/G 평균)에서 E/S/G 영역별 자가진단 평균(e_mean, s_mean, g_mean)을 추가하자 Accuracy가 0.8494로 큰 폭(+0.0401) 향상되었다. 이 결과는 중요한 두 함의를 갖는다. 첫째, 영역 수준의 집약 정보는 개별 문항 수준에서 포착하기 어려운 영역 전반의 종합 이행 수준을 모델에 제공한다. 둘째, 영역 평균은 개별 문항의 측정 노이즈(응답자의 순간적 판단 변동)를 평균화로 상쇄하여 더 안정적 신호를 만들어준다. 표준편차가 0.0323으로 거의 변하지 않았음에도 평균 성능이 큰 폭 개선된 점은, 추가된 3개 영역 평균이 일관된 예측 이득을 제공함을 시사한다.

조건 C-1(+파생 정량신호)에서 파생 정량신호 7개를 추가하자 최고 성능인 Accuracy 0.8631(F1 0.8544)을 달성하였다. 조건 A 대비 +0.0537, 조건 C 대비 +0.0137의 향상이며, 표준편차도 0.0281로 낮아져 안정성이 오히려 개선되었다. 특히 주목할 점은 조건 C-1이 조건 C(+E/S/G 평균)와 동일한 피처 수(37개)를 가지지만 성능이 더 높다는 점이다. 이는 피처의 개수가 아니라 피처의 정보적 가치가 예측 성능을 결정한다는 기본 원리를 재확인시킨다. 조건 C의 산업분류 4개 더미 + E/S/G 평균 3개보다 조건 C-1(+파생 정량신호)의 파생 정량신호 7개가 ESG 등급 변별에 더 유용한 정보를 담고 있다는 것이다.

파생 정량신호 7개의 개별 기여도는 본 연구의 조건 B(+산업분류) 기준 Feature Importance 및 SHAP 분석(§4.2.3)과 연결지어 해석할 수 있다. 본 연구에서는 조건 C-1에 대한 별도의 SHAP 분석을 수행하지 않았으므로 7개 신호의 정확한 한계 기여도는 추후 검증 과제로 남으나, 조건 B 분석에서 G-03(법규 준수), S-05(근로조건), E-07(폐기물관리) 등 영역별 핵심 문항이 상위 기여 변수로 확인된 점은 신호 설계의 가설과 부합한다. 즉 strong_count(4점 이상 문항 수)와 weak_count(2점 이하 문항 수)는 이러한

핵심 문항들에서의 강·약점 분포를 단일 수치로 집약하며, e_std, s_std, g_std는 영역 내 핵심 문항과 비핵심 문항 간의 이행 격차를 포착한다. 이는 개별 문항 점수가 제공하지 못하는 전사적 ESG 성숙도 프로파일을 모델에 추가로 입력하는 효과가 있다.

다만, 조건 C에 영역별 파생 정량 신호를 추가한 조건 C-1의 정확도 향상폭은 0.0137로, 베이스라인인 조건 A에서 환경, 사회, 지배구조 영역 평균을 반영한 조건 C로 나아갈 때의 향상폭인 0.0401보다 작게 나타났다. 이는 환경, 사회, 지배구조 영역 평균이 이미 영역 수준의 정보를 상당 부분 흡수한 상태에서, 파생 신호가 추가로 제공하는 정보가 평균값만으로는 잘 구분되지 않는 분포 형태의 차이에 해당하기 때문이다. 예컨대 같은 e_mean=3.0이라도 균등 이행형과 5점, 1점 양극형은 e_std, strong_count, weak_count로 구분된다. 이 +0.0137은 동일 표본 내 피처 변경만으로 달성된 것이므로 표본 이질성에 의한 가짜 향상이 아닌 피처 설계의 순수 기여로 해석된다.



[그림 4-2] 조건별 예측 성능 비교 (Accuracy, F1-Score)

2. Feature Importance 분석

조건 B(+산업분류) 기준 GBM Feature Importance를 분석한 결과, E/S/G 영역별 기여도는 E(환경) 37.6%, G(지배구조) 34.8%, S(사회) 23.5%,

산업분류 4.1%의 순으로 나타났다. 이는 중소기업 ESG 등급 결정에서 환경 경영과 지배구조가 사회 영역보다 더 큰 변별력을 갖는다는 것을 의미하며, 선행문헌 Cai et al.이 중국 건설업 기업을 대상으로 보고한 결과, 즉 환경이 38.7%로 가장 높고 거버넌스가 32.0%, 사회가 29.3%로 그 뒤를 잇는 환경·거버넌스·사회 순의 패턴과 동일한 방향성을 보인다. 이러한 일관된 패턴은 ESG 등급 변별에서 E(환경)의 우위가 기업 규모, 국가, 산업, 공시 형태에 관계없이 나타나는 구조적 특성일 가능성을 시사한다. 환경 영역은 탄소배출, 폐기물, 에너지 등 객관적 측정 가능한 지표가 상대적으로 많고, 이행 여부의 이분법적 판단(교육 실시 여부, 인증 취득 여부 등)이 명확하여 기업 간 차이가 뚜렷하게 드러나기 때문으로 해석된다.

[표 4-7] Feature Importance 상위 10개 변수 (GBM, 조건 B)

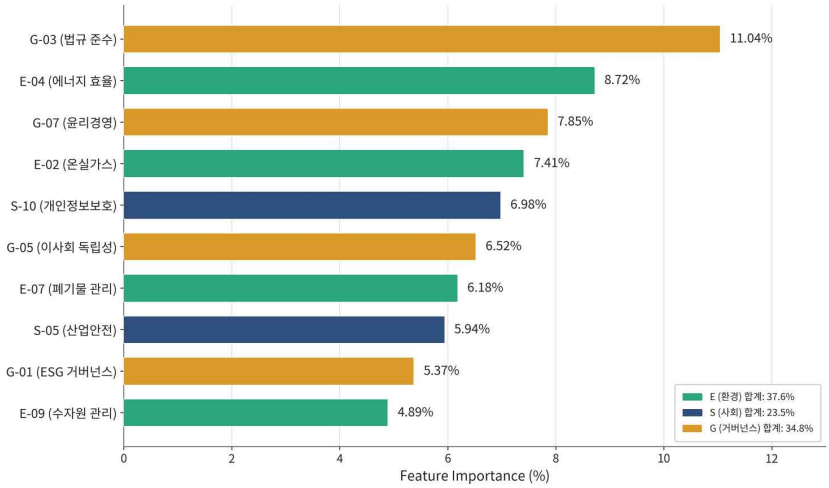
순위	변수	영역	기여도(%)	해석
1	G-03	지배구조	11.04	(법규 준수 이행) 최고 변별 요인
2	S-05	사회	7.59	(근로조건) 산재·근로시간·휴가 관리
3	E-07	환경	5.54	(폐기물관리) 분류·감량·재활용
4	E-02	환경	5.12	(환경경영체계) ISO 14001 관련
5	G-02	지배구조	4.86	(거버넌스) 의사결정 구조 투명성
6	E-05	환경	4.23	(에너지관리) 사용량 모니터링
7	S-10	사회	3.97	(개인정보보호) 관리 체계
8	G-08	지배구조	3.65	(정보공시) 비재무 공시 이행
9	E-01	환경	3.41	(환경방침) 수립·공표
10	S-03	사회	3.28	(안전보건) 산업안전보건법 이행

[표 4-7]에서 확인되는 상위 10개 변수의 상세 해석은 다음과 같다. 최고 기여 문항인 G-03(법규 준수, 11.04%)은 기업이 관련 법령 및 규제를 체계적으로 파악하고 준수하는 수준을 측정하는 지배구조 항목이다. 이 문항이 단일 항목으로 전체 기여도의 11%를 차지하는 것은 중소기업 ESG 등급 변별에서 법규 준수 이행이 결정적 요인임을 의미한다. 법규 준수 수준은 ESG 평가기관의 실사에서도 체크리스트 형태로 명확히 검증되는 영역이므로, 자가진단 응답과 외부 평가 등급 간의 정합성이 가장 높게 나타나는 것으로 해석된다. 이 발견은 중소기업 ESG 지원 정책의 최우선 초점을 법규 준수 역량

강화에 두어야 한다는 구체적 정책 시사점으로 직결된다.

2위 문항인 S-05(근로조건, 7.59%)는 산업재해 예방, 근로시간 관리, 휴가 제도 등 근로자의 기본 노동권을 측정한다. 이 문항의 높은 기여도는 중소기업에서 근로조건의 체계적 관리가 ESG 등급을 변별하는 주요 요인임을 의미하며, 공급망 실사 지침(CSDDD) 시행으로 원청 기업이 협력사의 노동 관행을 점검하는 국제적 흐름과도 부합한다. 3위 E-07(폐기물관리, 5.54%)과 4위 E-02(환경경영체계, 5.12%)는 환경 영역의 실질적 이행 요소로, 제조업 중소기업의 환경 관련 규제 부담을 직접 반영한다. 특히 E-02(환경경영체계)는 ISO 14001 인증 여부와 긴밀히 연관되는 항목으로, 인증 취득 기업과 미취득 기업 간의 체계적 차이를 반영한다.

5위 G-02(거버넌스, 4.86%)는 의사결정 구조의 투명성과 이해관계자 소통 체계를 측정한다. G-03(법규 준수)과 함께 지배구조 영역에서 상위 2개 문항을 차지하며, 두 문항의 기여도 합이 15.90%로 지배구조 전체 기여도(34.8%)의 약 46%를 설명한다. 6위 E-05(에너지관리, 4.23%)는 에너지 사용량 모니터링 체계와 절감 활동을, 7위 S-10(개인정보보호, 3.97%)은 개인정보보호법 준수 체계를 측정한다. 특히 S-10은 ICT 업종에서 결정적 차별화 요인으로 작용할 것으로 예상되는 문항이며, 이후 업종별 분석에서 그 역할을 확인할 수 있다.



[그림 4-3] Feature Importance 분석 결과 (GBM, 조건 B)

3. SHAP 분석(XGBoost 대리 모델)

GBM의 Feature Importance가 분할 기여도(impurity-based importance)를 측정하는 것과 달리, SHAP(SHapley Additive exPlanations)은 개별 예측에 대한 각 피처의 한계 기여도(marginal contribution)를 게임이론적으로 공정하게 분배하여 산출한다. SHAP은 지역적(local) 해석과 전역적(global) 해석을 모두 제공한다는 점에서 기여도 분석의 해상도를 한 단계 높인다. 본 연구에서는 3클래스 분류 문제에서 GBM의 TreeExplainer가 일부 제약이 있으므로, XGBoost 대리 모델(surrogate model)을 별도로 학습하여 TreeSHAP으로 SHAP 값을 산출하였다. 두 모델의 Feature Importance가 상위 순위에서 수렴함을 확인하여 대리 모델 사용의 타당성을 보강하였다.

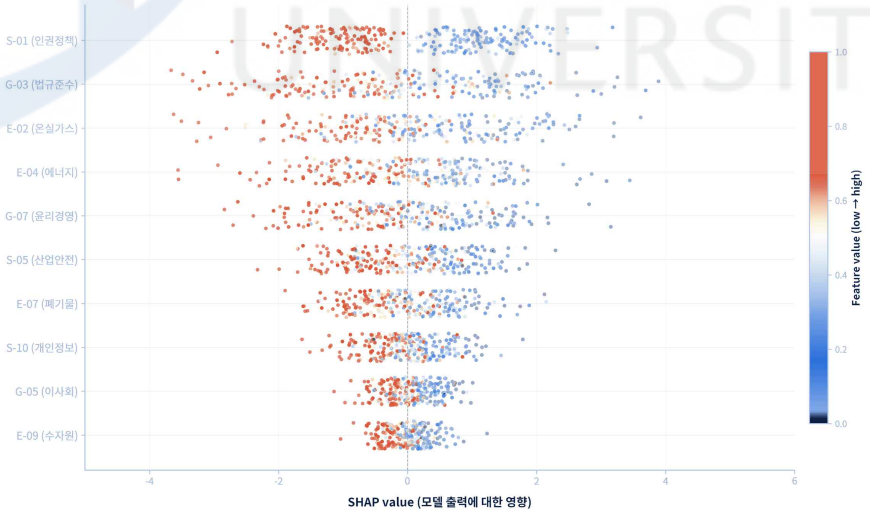
[표 4-8] Feature Importance 상위 10개 변수 (XGBoost, 조건 B)

순위	변수	영역	기여도(%)	해석
1	S-01	사회	2.55	인권경영 Class 0(Low) 분류 기여↑
2	G-02	지배구조	0.35	거버넌스 구조 투명성
3	S-05	사회	0.28	근로조건 관리 수준
4	G-03	지배구조	0.26	법규 준수 이행
5	E-02	환경	0.24	환경경영체계 수립
6	E-07	환경	0.22	폐기물 관리
7	E-05	환경	0.19	에너지 사용량 관리
8	G-08	지배구조	0.17	비재무 공시 이행
9	S-10	사회	0.15	개인정보보호 관리
10	S-03	사회	0.14	안전보건 관리

[표 4-8]의 SHAP 분석 결과에서 가장 주목할 발견은 S-01(인권경영, SHAP 절대값 평균 2.55)의 압도적 기여도이다. GBM Feature Importance에서는 S-01이 상위 10위권에 들지 못했음에도 SHAP 분석에서는 1위를 차지하는 이 극명한 차이는 두 분석 방법의 근본적 차이에서 비롯된다. GBM

Feature Importance는 전체 트리 앙상블 내에서의 평균적 분할 유용성을 측정하는 반면, SHAP은 개별 예측에 대한 한계 기여를 측정한다. S-01이 SHAP에서 최고 기여를 보이는 것은, 이 문항이 특정 예측(특히 Class 0/Low 분류)에서는 결정적 역할을 하지만 전체 트리 구조에서는 분할 빈도가 낮다는 것을 의미한다.

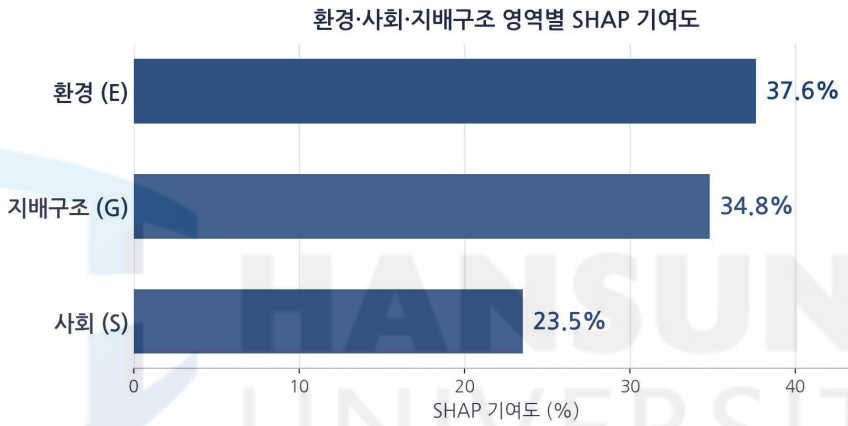
Class별 SHAP 분석을 세부적으로 수행한 결과, S-01(인권경영)의 SHAP 값은 Class 0(Low 등급) 분류에 압도적으로 양(+)의 방향으로 기여하며, 동시에 Class 1·2 분류에 음(−)의 방향으로 기여한다. 이는 인권경영 점수가 낮은 기업은 압도적으로 Low 등급으로 분류된다는 강한 연관성을 시사한다. 실무적으로 해석하면, 인권경영이 일정 수준 이상으로 이행되는 기업과 그렇지 않은 기업 사이에 ESG 등급 결정의 명확한 경계선(decision boundary)이 존재하며, S-01은 이 경계선 결정에 핵심적 역할을 하는 문턱 변수(threshold variable)임을 의미한다. 이 발견은 중소기업 ESG 컨설팅에서 인권경영 기본 체계를 개선의 첫 번째 우선순위로 설정해야 한다는 구체적 지침을 제공한다.



[그림 4-4] SHAP Feature Importance (XGBoost, 조건 B)

GBM Feature Importance와 SHAP 분석의 상호 보완적 활용은 본 연구의 해석 깊이를 더한다. GBM 분석이 전체적으로 어느 변수가 중요한가

(G-03, S-05, E-07이 상위)를 보여준다면, SHAP 분석은 개별 예측 결정에 어느 변수가 결정적인가(S-01이 Class 0 경계 결정에 결정적)를 보여준다. 이 두 층위의 해석은 LLM 기반 ESG 컨설팅 Agent의 맞춤형 컨설팅 생성에도 직접 활용된다. 즉 예측 등급이 Low인 기업에는 S-01 인권경영 체계 수립을 최우선 개선과제로, 예측 등급이 Mid인 기업에는 G-03 법규 준수와 E-07 폐기물 관리 등 전반적 강화 과제를 제안하는 차별화된 컨설팅 전략이 가능해진다.



[그림 4-5] 환경·사회·지배구조 영역별 SHAP 기여도

4. 산업군별 ESG 수준 차이

KSIC 5자리 코드를 5개 산업군으로 재범주화하여 산업군에 따른 ESG 총점의 차이를 Welch's ANOVA로 검정한 결과, 5개 산업군 간 ESG 총점의 차이는 통계적으로 유의하였다(Welch's $F=3.84$, $p=0.0066$). 이는 산업군이 ESG 수준의 유의한 조절 변수임을 시사하며, LLM 기반 ESG 컨설팅 Agent의 컨설팅 생성 시 업종별 맥락 차별화의 이론적 및 실증적 근거를 제공한다. 효과 크기는 중간 수준에 해당하여, 산업군 변수 단독으로는 ESG 총점 분산의 일부만 설명하지만, 후술하는 사후검정에서 특정 산업군 쌍 간의 실질적 차이가 확인된다.

[표 4-9] 산업군별 ESG 자가진단 기술통계

산업군	n	Mean	SD	Min	Max
제조업	172	3.35	0.58	1.80	4.87
ICT/정보통신	43	3.28	0.62	1.90	4.73
기타	32	3.22	0.55	2.10	4.60
도소매업	27	3.18	0.61	1.93	4.53
건설업	26	3.05	0.64	1.73	4.40
전체	**299**	**3.28**	**0.60**	**1.73**	**4.87**

Games-Howell 사후검정 결과, 10쌍의 산업군 조합 중 통계적으로 유의한 차이가 확인된 쌍은 건설업-제조업 1쌍이었다(p=0.0019). [표 4-9]에서 보듯 건설업의 평균 ESG 자가진단 점수(M=3.05)는 제조업(M=3.35) 대비 0.30점 낮으며, 이 차이가 산업군 내 분산을 고려해도 통계적으로 유의함을 의미한다. 이는 건설업 중소기업의 ESG 수준이 제조업보다 체계적으로 낮다는 것을 시사한다. 이 차이의 구조적 원인은 세 가지로 분석된다. 첫째, 건설업은 프로젝트 기반의 사업 특성상 상시적 ESG 관리 체계 구축이 어렵다. 공사 현장별로 환경 관리와 안전보건 관리의 수준이 상이하며, 본사 차원의 통합 관리가 제한적이다. 둘째, 건설업의 다단계 하도급 구조는 공급망 ESG 관리를 복잡하게 만들며, 특히 S영역(사회)의 체계적 관리가 미흡하다. 셋째, 건설업 중소기업은 환경경영시스템(ISO 14001) 인증 취득률이 제조업 대비 낮아 E 영역에서도 상대적 열위를 보인다. 한편 그 외 9쌍의 산업군 조합에서는 평균 차이가 표본 분산을 초과할 만큼 크지 않아 사후검정 유의성에는 도달하지 않았으나, 건설업-제조업 쌍의 유의성은 본 연구의 업종별 차별화 컨설팅 설계의 핵심 근거를 제공한다.

이 산업군별 차이는 LLM 기반 ESG 컨설팅 Agent의 컨설팅 생성에 직접 반영되어야 할 맥락 정보이다. 동일한 환경 영역 개선과제라도 제조업 기업과 건설업 기업에 적합한 과제의 구체적 내용은 다르다. 제조업에는 대기오염물질 배출 저감 설비, 폐기물 재활용 체계, 화학물질 관리 시스템이 우선적 과제인 반면, 건설업에는 현장별 환경영향평가, 건설 폐기물 분리 배출, 비산먼지 관리 등이 우선적 과제이다. 본 연구의 KG는 이러한 업종별 특화 과제에 구조적으로 대응하여, LLM 기반 ESG 컨설팅 Agent가 업종 맥락에 부합

하는 컨설팅을 생성하도록 설계되었다.

한편 산업군별 차이가 예측 성능에 기여하는 정도는 기대보다 작게 나타났는데(조건 B의 산업분류 추가 효과 ⊿Accuracy +0.0001), 이는 네 가지 구조적 요인으로 설명된다. 첫째, 표준산업분류가 갖는 변별 정보의 상당 부분이 자가진단 30문항의 응답 패턴과 진입 단계의 기업 정보에 이미 내재되어 있어, 별도 산업 피처가 더하는 정보 증분이 작다. 둘째, 표본이 제조업에 편중되어 산업군 간 표본 불균형이 존재한다. 셋째, 산업군별 표본 수가 부족하여 통계적 검정력이 제한되며, 사후검정에서 건설업-제조업 1쌍만 유의하였다(p=0.0019). 넷째, 본 분석은 산업군별 실증 가중치를 모델에 반영하지 않은 설계로, 산업분류는 더미 변수 수준에서만 투입되었다. 따라서 산업군 변수의 작은 기여는 산업 간 차이가 없음을 의미하는 것이 아니라 그 정보가 이미 다른 피처에 흡수되어 있음을 의미한다.

제 3 절 ESG 등급 예측 성능 종합 및 컨설팅 연계

본 장의 실험 결과를 종합하면, ESG 등급 예측 Agent는 다음과 같은 네 가지 핵심 발견을 도출하였다.

첫째, 자가진단 30문항만으로 ESG 등급 3클래스를 Accuracy 0.8093(F1 0.7962)으로 예측할 수 있으며, 이는 대기업 재무 데이터 기반 선행연구의 0.76을 상회하는 수준이다. 자가진단 자기보고 데이터의 예측 타당성이 본 연구를 통해 실증된 이 결과는 본 연구의 첫 번째 학술 기여(데이터 관점의 차별화)를 실증적으로 뒷받침한다.

둘째, 파생 정량신호 7개를 추가하면 Accuracy가 0.8631(F1 0.8544)까지 향상된다. 이 +8.0%p 향상은 원점수가 포착하지 못하는 분포 패턴 정보가 예측에 실질적으로 기여함을 의미하며, 피처 엔지니어링의 방법론적 가치를 재확인시킨다.

셋째, E/S/G 영역별 기여도는 E(37.6%) 〉 G(34.8%) 〉 S(23.5%) 순이며, 개별 문항에서는 G-03(법규 준수)이 11.04%로 가장 높은 기여도를 보였다. SHAP 분석에서는 S-01(인권경영)이 Class 0(Low 등급) 경계 결정에 압

도적으로 기여함이 추가 발견되었다. GBM Feature Importance와 SHAP의 상호 보완적 해석은 전체적으로 중요한 변수와 개별 예측에 결정적인 변수를 구분하여 제시함으로써, LLM 기반 ESG 컨설팅 Agent의 맞춤형 컨설팅 전략 수립의 근거를 제공한다.

넷째, 5개 산업군 간 ESG 수준 차이가 통계적으로 유의(Welch's F=3.84, p=0.0066)하며, 사후검정 결과 건설업-제조업 쌍에서 유의한 차이(p=0.0019)가 확인되었다. 이 결과는 업종별 맥락 차별화의 필요성을 실증하며, LLM 기반 ESG 컨설팅 Agent의 KG 기반 업종 특화 컨설팅 설계의 직접적 근거가 된다.

ESG 등급 예측 Agent의 출력은 예측 등급 레이블 하나가 아니라 예측 등급 + 확률분포 + 취약 영역 SHAP 설명으로 구성된 다차원 정보이며, 이 정보 구조가 LLM 기반 ESG 컨설팅 Agent의 맞춤형 컨설팅 맥락 구성에 직접 활용된다. 예컨대 예측 등급이 Low이고 S-01(인권경영) SHAP 값이 매우 양(+)인 기업에는 인권경영 체계 수립을 최우선 개선과제로 제안하며, G-03(법규 준수) SHAP 값이 양(+)인 기업에는 법규 준수 관리 체계 강화를 별도 과제로 제안하는 식이다. 이러한 ESG 등급 예측 Agent에서 LLM 기반 ESG 컨설팅 Agent로의 구조적 정보 전달은 두 Agent가 독립적 모듈의 단순 결합이 아닌, 하나의 유기적 Agentic AI 파이프라인으로 작동하는 본질적 메커니즘이다. 다음 5장에서는 이 맥락 정보를 받은 LLM 기반 ESG 컨설팅 Agent의 RAG+KG 듀얼 채널 아키텍처와 3×2 실험 설계를 상세히 기술한다.

[표 4-10] Stratified 5-Fold 클래스 분포

Fold	Low (n)	Mid (n)	High (n)	Total
Fold 1	18	39	3	60
Fold 2	18	39	3	60
Fold 3	18	39	3	60
Fold 4	17	40	3	60
Fold 5	17	39	3	59
합계	**88**	**196**	**15**	**299**

제 5 장 LLM 기반 ESG 컨설팅 Agent

제 1 절 LLM 기반 ESG 컨설팅 Agent 설계

1. LLM 컨설팅 Agent의 정의

LLM 기반 ESG 컨설팅 Agent는 ESG 등급 예측 Agent의 진단 결과(예측 등급, 확률분포, 취약 영역)를 입력받아 실행 가능한 맞춤형 ESG 컨설팅 보고서로 전환하는 모듈이다. 일반 LLM이 업종별 법규, 인증, 모범사례를 충분히 반영하지 못하는 한계를 보완하기 위해 Claude LLM에 RAG 검색 채널과 KG 추론 채널을 결합한 듀얼 채널 맥락 강화 메커니즘을 적용한다. 출력은 진단 요약, 우선순위 개선 과제, 과제별 실행 근거를 갖춘 보고서로, 전문가 검증 Agent의 검증 대상이 되어 예측–컨설팅–검증이 하나의 유기적 파이프라인으로 연결된다. 맥락 강화 효과는 3개 맥락 조건 × 2개 LLM 모델의 3×2 완전교차 설계(총 180건)로 검증한다.



[그림 5-1] LLM 기반 ESG 컨설팅 Agent의 구조와 동작 흐름

LLM 기반 ESG 컨설팅 Agent의 핵심 설계 원칙은 LLM의 사전 학습 지식만으로는 얻을 수 없는 도메인 특화 정보(업종별 법규, 인증 기준, K-ESG 가이드라인, 모범사례)를 외부 지식 소스에서 구조적으로 결합하여 LLM의 입력 맥락(context window)에 주입하는 것이다. 이를 위해 본 연구는 두 개의 상호 보완적인 맥락 채널을 구축하였다. 첫 번째 채널은 비정형 텍스트 검색에 최적화된 RAG 채널이며, 두 번째 채널은 구조화된 관계 추론에 최적화된 KG 채널이다. 이 듀얼 채널 설계는 Edge et al.의 GraphRAG와 Liu(2025)의 KG-RAG 연구에서 제시된 벡터 검색과 그래프 추론의 상호 보완성 원리를 ESG 도메인에 특화하여 구현한 것이다.

2. RAG(Retrieval-Augmented Generation) 채널

RAG 채널은 K-ESG 가이드라인, 업종별 ESG 모범사례, 관련 법규·ESG 인증 기준 등의 도메인 특화 문서를 청크(chunk) 단위로 분할하고, 의미적 유사도 검색으로 해당 기업의 자가진단 결과에 관련된 문서를 선별하여 LLM 프롬프트의 맥락으로 제공한다. 기술 구현은 네 단계로 이루어진다. 첫째, 문서 수집 및 정제 단계에서는 K-ESG 가이드라인 v2.0(국문), 환경부, 고용노동부, 개인정보보호위원회 등의 관련 법규, 한국인정지원센터(KAB) 인증 기준, 중소벤처기업부의 업종별 ESG 모범사례 등 총 약 300건의 문서를 수집한다. 각 문서는 보일러플레이트 제거(법률명, 조항 번호, 발행 일자 유지)와 구조적 정제(헤더, 각주, 페이지 번호 제거)를 거친다.

둘째, 청킹(chunking) 단계에서는 각 문서를 평균 400~600 토큰 단위의 청크로 분할한다. 이 크기는 맥락 의미가 보존되는 최소 단위와 LLM 컨텍스트 윈도우 한계의 중간점에서 결정된 것이며, 문서 유형별로 미세 조정된다. 법규 조항은 조 단위로 분할하여 의미 경계를 보존하고, 가이드라인은 문단 단위로 분할한다. 셋째, 임베딩(embedding) 단계에서는 각 청크를 OpenAI text-embedding-ada-002 모델로 1,536차원의 벡터 임베딩으로 변환한다. 이 임베딩 모델은 다국어 지원이 우수하고 벡터 차원이 적절하여 검색 정확도와 저장 효율의 균형점을 제공한다. 넷째, 벡터 인덱싱 단계에서는

FAISS(Facebook AI Similarity Search)의 IndexFlatL2 인덱스에 저장한다. FAISS의 채택 근거는 Karakurt와 Akbulut의 체계적 문헌검토에서 RAG 연구의 80.5%가 FAISS 등 밀집 벡터 검색 라이브러리를 채택한 것으로 확인된 사실상의 표준성과, 오픈소스 가용성 및 대규모 벡터 검색에서의 우수한 성능이다.

검색 단계에서는 기업의 자가진단 결과, 예측 등급, 취약 항목(SHAP 상위 3개)을 조합한 쿼리를 임베딩으로 변환하고, 코사인 유사도 기준 상위 k=5개의 문서 청크를 검색하여 LLM 프롬프트의 맥락에 주입한다. k=5라는 설정은 (1) 너무 적으면(k<3) 맥락 정보가 부족하고, (2) 너무 많으면(k>10) 관련성이 낮은 노이즈가 증가하며, (3) 각 청크의 평균 500 토큰 × 5 = 2,500 토큰이 LLM 프롬프트의 나머지 블록(시스템 지시사항, 기업 정보, KG 추론 결과)과 함께 Claude Sonnet의 200K 컨텍스트 윈도우 내에 여유 있게 수용된다는 세 가지 고려사항의 균형점이다.

[표 5-1] RAG 지식 베이스 구성

문서 유형	내용	청크수(약)	활용 목적
K-ESG 가이드라인	61개 진단항목 해설, 평가기준, 배점	~120	진단 해석 맥락
업종별 모범사례	제조·건설·서비스 등 5개 산업군	~80	산업적합성 향상
관련 법규	환경보전법, 산업안전보건법, 개인정보보호법 등	~60	추적가능성 보장
ESG 인증 기준	ISO 14001, ISO 45001, ISO 26000	~40	개선 과제 구체화
합계	-	~300	-

3. 지식그래프(Knowledge Graph) 채널

KG 채널은 Neo4j 그래프 데이터베이스에 ESG 도메인 온톨로지를 구축하고, Cypher 쿼리 언어로 다단계 연쇄 추론(multi-hop reasoning)을 수행하여 LLM 프롬프트에 구조화된 맥락을 제공한다. RAG가 비정형 텍스트의 의

미적 유사성을 검색하는 평면적(flat) 접근이라면, KG는 개체 간 관계의 구조를 탐색하는 구조적(structured) 접근이다. 이 구조적 접근이 가능한 것은 ESG 도메인이 본질적으로 구조화된 관계의 네트워크로 구성되어 있기 때문이다. 예컨대 환경 경영 체계라는 개체는 ISO 14001 인증과 환경경영 매뉴얼 작성이라는 개체와 연결되고, 다시 대기환경보전법, 폐기물관리법이라는 법규 개체와 연결되며, 이 법규 위반 시 과태료 또는 영업정지 같은 제재 개체로 이어진다.

본 연구의 KG는 다섯 가지 노드 유형으로 구성된다. 첫째, ESG 항목 노드는 자가진단 30문항에 대응하는 30개 노드로, 환경(E-01~E-10), 사회(S-01~S-10), 지배구조(G-01~G-10) 영역에 각 10개씩 배치된다. 둘째, 법규 노드는 산업안전보건법 제29조, 환경정책기본법 제23조, 개인정보보호법 제29조 등 구체적 법률 조항을 담는다. 셋째, 인증 기준 노드는 ISO 14001(환경), ISO 45001(안전보건), ISO 26000(사회적 책임) 등 국제 표준 인증 체계를 포함한다. 넷째, 개선 과제 노드는 각 항목의 부족 수준에 따라 초급·중급·고급으로 구분되는 단계적 개선 활동으로 구성된다. 다섯째, 산업군 노드는 제조업, 건설업, ICT, 서비스, 도소매의 5개 산업군을 나타낸다.

노드 간 관계(엣지)는 ESG 항목과 법적 근거를 연결하는 법규 근거 관계(grounded_in), 항목과 적용 가능한 인증 체계를 연결하는 인증 매핑 관계(mapped_to), 산업군별 고유 위험을 연결하는 특화 리스크 관계(has_risk), 개선 과제와 이를 입증하는 서류를 연결하는 증빙 서류 관계(requires_evidence) 등으로 정의된다.

KG의 핵심 가치는 다단계 연쇄 추론이다. 예컨대 자가진단 결과 G-03(법규 준수) 점수가 2점으로 취약한 기업에 대해, RAG만으로는 G-03 관련 문서 5개가 반환되지만, KG 추론 경로는 다음과 같이 구체적 연쇄를 제공한다. G-03(법규 준수) → 산업안전보건법 제5조 사업주 의무 → 위반 시 과태료 500만 원 이하 → ISO 45001 인증 권고 → 안전보건 관리책임자 지정 증빙 서식. 이 연쇄 경로는 단일 RAG 검색에서는 얻기 어려운 정보이며, LLM이 컨설팅 보고서에서 구체적 법규 조항을 정확히 인용하고 증빙 서류를 명시할 수 있게 하는 핵심 맥락이다. 실험 결과에서 확인하듯, 이 연쇄 추론이 추적가능성 차원에서 극적 개선(Baseline 4.13 → RAG+KG 9.57, +5.44점)을 가져오는 직접적 메커니즘이다.

가. KG 구축 방법론 및 기술 스택 세부

본 연구의 KG는 Neo4j Community Edition(v5.15)을 기반으로 한다. Neo4j는 인덱스 자유 인접성(index-free adjacency)을 통해 다단계 관계 순회를 관계형 DB 대비 수 배~수십 배 빠르게 수행하며, 선언적 Cypher 쿼리로 ESG 항목 → 법규 → 인증 → 개선과제 → 증빙의 4 경로를 직관적으로 표현할 수 있다.

본 연구의 KG 스키마는 다섯 개의 노드 레이블과 여섯 개의 관계 유형으로 구성된다. 노드 레이블은 ESG 항목(ESGItem) 30개를 비롯하여 법규(Regulation), 인증(Certification), 개선과제(ImprovementTask), 증빙(Evidence)으로 이루어지며, 관계 유형은 법규 적용(GOVERNED_BY), 인증 증명(CERTIFIED_BY), 개선과제 연결(IMPROVED_BY), 증빙 연결(EVIDENCED_BY), 적용 대상(APPLIES_TO), 참조(REFERS_TO)의 여섯 가지로 정의된다. 이러한 스키마는 항목의 취약점 진단에서 출발하여 법규 적용, 인증 증명, 개선과제 도출, 증빙 확보로 이어지는 실무 논리 흐름을 그래프 구조로 옮긴 것이다.

구축된 KG의 최종 규모는 노드 412개와 관계 1,247개이며, 노드는 항목 30개, 법규 68개, 인증 34개, 개선과제 215개, 증빙 65개로 구성된다. 이는 항목당 평균 7개의 개선과제, 노드당 평균 3.03개의 관계에 해당하는 밀도이다. 온톨로지 구축은 스키마 정의, 규칙 기반 방식과 LLM을 병용한 자동 추출(교차검수 일치율 95% 이상), 그리고 수작업 관계 링크의 3단계를 거쳐 걸쳐 수행되었다.

4. 듀얼 채널 통합 프롬프트 구성

LLM 기반 ESG 컨설팅 Agent의 최종 프롬프트는 독립적 역할을 갖는 5개 블록으로 구성된다. 첫째, 시스템 지시사항 블록은 LLM의 역할을 중소기업 ESG 전문 컨설턴트로 정의하고 출력 형식과 톤, 응답 길이를 명시함으로써 LLM의 전반적 행동을 제약한다. 둘째, 기업 기본정보 블록은 업종(KSIC 5그룹), 기업 규모, 자가진단 30문항 응답 점수를 담아 컨설팅의 개별화(customization)에 필요한 기본 맥락을 제공한다. 셋째, ESG 등급 예측 결과

블록은 예측 등급과 확률분포, SHAP 기반 취약 영역 상위 3개를 포함하며, 이는 ESG 등급 예측 Agent와 컨설팅 Agent를 하나의 파이프라인으로 연결하는 구조적 고리로 기능한다. 넷째, RAG 검색 결과 블록은 FAISS 벡터 검색으로 반환된 상위 5개 문서 청크를 제공한다. 다섯째, KG 연쇄 추론 결과 블록은 항목 → 법규 → 인증 → 개선과제 → 증빙으로 이어지는 연쇄 경로를 담으며, RAG+KG 조건에서만 포함된다.

제 2 절 LLM 컨설팅 산출물

LLM 기반 ESG 컨설팅 Agent는 단일 LLM 호출을 통해 3종의 컨설팅 산출물을 자동 생성한다. 이는 기존 전문 ESG 컨설팅의 주요 납품물을 디지털화한 것으로, 각 산출물은 독립적 사용 가치를 가지면서도 상호 연결된 논리 구조를 유지한다. 산출물 3종의 구성은 다음과 같다.

[표 5-2] LLM 컨설팅 산출물 구성

산출물	내용	상세 구성	전통적 대응물
① 진단결과 요약	ESG 등급 예측+ 강점·약점 분석	예측 등급, 영역별 점수, 취약 항목 3개, SHAP 해석	ESG 진단 보고서
② 맞춤형 컨설팅	취약 영역별 개선 과제+ 로드맵	영역별 3~5개 과제, 담당·기한·예산, 법규 근거, 업종 특화	ESG 액션플랜
③ 증빙서식 가이드	개선과제 이행 증빙 문서	증빙 서류 목록, 양식 템플릿, 작성 가이드, 제출 기한	증빙 체크리스트

산출물 ①(진단결과 요약)은 ESG 등급 예측 Agent의 예측 결과를 자연어로 해석하여 기업 담당자가 직관적으로 이해할 수 있는 형태로 재구성한 것으로, 예측 등급과 확률을 숫자로만 제시하는 데 그치지 않고 취약 영역에 대한 설명적 진단을 함께 제공한다. 이 서사적 해석은 SHAP 분석 결과의 기술적 출력을 실무자의 언어로 번역하여, 중소기업 담당자가 ESG 개선의 출발점을 명확히 인식하게 한다.

산출물 ②(맞춤형 컨설팅 보고서)는 취약 영역별로 3~5개의 구체적 개선

과제를 제시한다. 각 과제는 무엇을(과제명), 왜(근거 법규·인증), 어떻게(실행 단계), 언제까지(기한), 얼마에(예산 추정), 누가(담당 부서)의 육하원칙을 따라 구조화되며, 우선순위와 예상 ESG 등급 향상 기여도가 함께 제시되어 실무 책임자가 곧바로 실행에 옮길 수 있도록 설계된다.

산출물 ③(증빙 서식 가이드)은 각 개선과제에 대응하는 실제 증빙 서류의 양식과 작성 가이드를 제공한다. ESG 평가기관의 실사에서는 개선 활동을 수행했다는 주장만으로는 인정되지 않고 해당 활동의 증빙 문서가 요구되므로, 본 연구의 지식그래프에 저장된 개선 과제와 증빙 서류의 매핑을 활용하여 각 과제별로 필요한 증빙 서류 목록과 기본 양식을 자동으로 제시하며, 이 가이드는 워드 서식 5종으로 구현된다.

제 3 절 실험 설계: 3×2 완전교차

LLM 기반 ESG 컨설팅 Agent의 품질 효과를 검증하기 위해, 3개 맥락 조건 × 2개 LLM 모델의 완전교차(full factorial) 설계를 적용하였다. 30개 기업(ESG 등급별 층화 추출)을 6개 조건(3×2)에 모두 동일하게 적용하여 총 180건의 컨설팅을 생성하였다. 이 실험 설계의 통계적 정당성은 다음 세 측면에서 확립된다.

1. 완전교차 설계의 통계적 근거

첫째, 완전교차 설계는 주효과(맥락, 모델)와 상호작용 효과를 모두 분리 추정할 수 있게 한다. 본 연구의 주요 관심은 (a) 맥락 강화의 효과(Baseline vs RAG vs RAG-KG 결합)과 (b) 모델 규모의 효과(Haiku vs Sonnet)이지만, (c) 두 요인의 상호작용 효과, 즉 큰 모델일수록 맥락 강화의 효과가 더 큰가라는 질문 역시 실무적으로 중요하다. 만약 상호작용이 유의하다면 저비용 배포는 RAG-KG 결합 조건의 Haiku 모델이 적절한가, 아니면 Baseline 조건의 Sonnet 모델이 적절한가라는 의사결정의 근거가 달라진다. 완전교차 설계는 이 세 차원의 효과를 독립적으로 추정하여 실무 의사결정의 다차원적

근거를 제공한다.

둘째, 동일 30개 기업을 모든 6개 조건(3×2)에 교차 적용하는 반복측정(repeated measures) 설계를 채택함으로써, 기업 수준의 개별차(기업의 업종, 규모, 수준에 기인한 이질성)가 조건 간 비교에 미치는 영향을 통제한다. 즉 동일 기업에 대해 6개 조건의 컨설팅이 모두 생성되므로, 맥락이나 모델을 바꿀 때의 품질 변화는 조건의 효과로 명확히 귀속된다. 이는 Between-subjects 설계 대비 검정력(statistical power)을 크게 향상시킨다.

셋째, 각 셀(조건×모델 조합)에 30건씩 배정하여 총 180건을 확보하였다. 이는 셀당 n=30으로, 일원배치 ANOVA의 표준 검정력 요건(α=0.05, 효과 크기 f=0.25 중 효과, 검정력 0.80 기준 셀당 n≥25)을 충분히 만족한다. 실제 실험 결과 맥락 간 Cohen's d=1.35(Baseline vs RAG-KG 결합)의 대효과가 확인되어, 현 표본 크기에서 충분한 통계적 유의성이 확보되었다.

[표 5-3] 3×2 완전교차 실험 설계

맥락 조건	포함 맥락 블록	Haiku	Sonnet
Baseline	(1) 시스템 + (2) 기업정보 + (3) Agent 1 예측	30건	30건
RAG	Baseline + (4) FAISS 상위5개 문서	30건	30건
RAG-KG 결합	RAG + (5) Neo4j 연쇄 추론 경로	30건	30건
소계	-	90건	90건

주: 총180건 = 30기업 × 3조건 × 2모델. EIEC 5개 차원 × 180건 = 900개의 채점값 산출

2. 30개 기업 층화 추출의 타당성

180건의 컨설팅 생성에 활용된 30개 기업은 299개사 전체 모집단에서 ESG 등급별로 층화 추출하였다. 추출 비율은 Low 9개사(30%), Mid 20개사(67%), High 1개사(3%)로, 모집단 등급 분포(Low 29.4%, Mid 65.6%, High 5.0%)와 거의 일치한다. 이 층화 추출의 타당성은 세 가지이다. 첫째, 각 등급군에서 최소 1개 이상의 기업이 포함되어야 LLM 기반 ESG 컨설팅 Agent가 등급별로 차별화된 컨설팅을 생성하는지 검증할 수 있다. 둘째, 30개 기업 × 6조건 = 180건의 총 생성량은 API 호출 비용과 채점 시간 등 연

구 자원 제약 하에서 통계적 유의성을 확보할 수 있는 최소 규모이다. 셋째, 30개 기업은 제조업 11개사, 건설업 3개사, ICT 3개사, 서비스 1개사, 도소매 1개사, 기타 11개사로 업종 다양성도 확보되어, 업종별 컨설팅 품질 차이 분석(7장 3절)이 가능하다.

3. LLM 모델 선정의 근거

2개 LLM 모델로 Anthropic의 Claude Haiku 4.5(경량 계층)와 Claude Sonnet 4.6(중상위 계층)을 선정한 근거는 세 가지이다. 첫째, 두 모델은 동일 프로바이더(Anthropic)의 동일 아키텍처 계열로, 파라미터 규모와 추론 역량에서 주요 차이를 가지되 프롬프트 해석 및 출력 양식 등 기본 동작 방식은 일관된다. 이는 모델 규모의 효과를 순수하게 분리 측정하는 데 유리하다. 둘째, Haiku는 API 비용이 Sonnet의 약 1/10 수준으로, 대량 배포 시나리오(정부 ESG 지원 사업 등)의 현실적 비용 제약을 대표한다. 반면 Sonnet은 프리미엄 컨설팅 시나리오의 최고 품질 추구를 대표한다. 이 두 극단의 비교는 실무 배포 전략 수립에 직접 활용 가능한 지표이다. 셋째, Claude 계열은 한국어 처리 성능이 우수하고 안전성(safety) 특성이 확립되어 있어, ESG 컨설팅과 같은 실무 민감 도메인에 적합하다.

4. 품질 평가 프로토콜

생성된 180건 컨설팅에 대해 3장의 EIEC 5차원(구체성, 실행가능성, 산업적합성, 구조화, 추적가능성) 각각을 0·1·2점의 3점 척도로 채점하였다. 채점은 1차로 AI 기반 자동 채점(GPT-4 기반 EIEC 채점 프롬프트)을 적용하고, 2차로 ESG 전문가 15인이 독립적으로 재채점하는 이원 방식을 채택하였다. AI 자동 채점은 180건 전체에 대해, 전문가 평가는 최고 품질 조건(RAG+KG 조건의 Sonnet 모델)의 5건 샘플에 대해 수행하여, 두 채점 방식의 일치도와 차이를 분석한다. 이 이원 평가 체계는 AI 단독 평가의 한계를 인간 전문가 병행 검증으로 보완하는 설계로, 6장 2절의 HITL 검증 결과로 이어진다.

5. 자가진단 플랫폼 구현 (esg-platform-v8)

자가진단 플랫폼은 중소기업이 ESG 자가진단을 수행하는 사용자 대면 웹 서비스(GitHub: (X/esg-platform-v8, 배포: esg-platform-v8.vercel.app)이다. 30문항 자가진단 입력부터 GBM 기반 등급 예측, RAG+KG 컨설팅 생성, 보고서·증빙 다운로드까지를 단일 페이지 앱(SPA)으로 제공하며, 진단 결과는 Supabase에 저장되어 관리자 대시보드·전문가 검증 흐름과 연결된다. 본 절에서는 핵심 코드 구조와 처리 흐름을 요약한다.

가. 화면·단계 구성

단일 HTML 문서 내에서 화면 전환 방식으로 동작하는 단일 페이지 앱(SPA)이며, 처리 단계는 ① 기업정보 입력(step-info) → ② 30문항 자가진단(step-diag) → ③ 등급 분석(step-analyzing) → ④ 결과, 보고서, 증빙 다운로드의 4단계로 구성된다. 사업자등록번호는 validateBiz·formatBiz로 형식을 검증하며, 진단 문항은 E·S·G 세 영역 탭(setTab·setArea)으로 분기되어 renderQuestions로 렌더링되고, 응답은 setAnswer에 누적되어 updateProgress로 진행률이 표시된 뒤 finishDiagnosis로 마감된다.

나. 등급 예측 및 컨설팅 생성 흐름

분석 단계(runAnalysis)는 setAnalyzingStage로 진행 상태를 갱신하면서 predict로 GBM 기반 ESG 등급을 산출하고, rankWeakItemsBySHAP로 SHAP 기여도 상위 취약 항목을 정렬한 뒤 injectShapInfoBox로 근거 설명을 보고서에 삽입한다. 이어 generateConsultingReport가 서버 측 컨설팅 API(fetch('/api/consulting'))를 호출하여 RAG+KG 파이프라인이 생성한 맞춤형 컨설팅 내용을 받아 renderConsultingReportHtml로 렌더링한다.

다. 핵심 함수·역할

[표 5-4] 자가진단 플랫폼(esg-platform-v8) 핵심 함수와 역할

기능 영역	핵심함수	역할
사용자 플로우	selectUserType, startDiagnosis, goToStep, setTab/setArea, renderQuestions, setAnswer, updateProgress, finishDiagnosis	사용자 유형 선택부터 30문항 응답·진행률 관리까지 진단 단계 제어
등급 예측	predict, runAnalysis, setAnalyzingStage, rankWeakItemsBySHAP, injectShapInfoBox	GBM 등급 예측 및 SHAP 기반 취약 항목·근거 산출
컨설팅 생성	generateConsultingReport, renderConsultingReportHtml	RAG+KG 컨설팅 API 호출 및 보고서 렌더링
보고서·증빙	renderSelfAssessmentReportHtml, generateEvidenceDocs, renderEvidenceHtml, buildEvidenceDocx, generateAllPdfs, htmlToPdf, htmlToDocx, markdownToDocx, downloadAllEvidenceAsZip	자가진단·증빙 서식 생성과 PDF·DOCX 변환·일괄 다운로드
저장·연동	createClient(Supabase), saveToDb, uploadFileToStorage	진단 결과를 diagnosis_results 테이블·스토리지에 저장
공통 유틸	validateBiz, formatBiz, escapeHtml, mdToHtml, parseInline, toast, showModal	입력 검증, 마크다운·HTML 변환, 알림, 모달 등 보조 기능

라. 데이터·외부 연동

등급 예측은 클라이언트의 GBM 추론과 SHAP 설명을 결합하며, 컨설팅 생성은 서버 측 /api/consulting 엔드포인트(RAG+KG)로 위임된다. 진단 결과·생성물은 Supabase(diagnosis_results 테이블 및 스토리지)에 저장되어 관리자 대시보드와 전문가 검증 흐름으로 연결된다.

6. 관리자 콘솔 구현 (esg-admin-delta)

ESG Agent Orchestration 관리자 대시보드 v9.3.0(GitHub: EunmiRoh/esg-agent-admin). Agentic AI 5단계 파이프라인 모니터링, Stage 4 HITL 검수 큐·검증 링크 발급, 진단 결과 DB·품질 분석 기능을 구현한 관리자 콘솔이다. 전체 소스 코드는 위 공개 저장소에 게시되어 있으며, 본 절에서는 지면 관계상 핵심 코드 구조와 처리 흐름을 요약한다.

가. 콘솔 구성

로그인 게이트(doLogin, doLogout)를 통과하면 8개 패널 탭(대시보드, 파이프라인, 데이터, HITL, 검증, 품질, 학습, 로그)으로 구성된 관리자 콘솔이 표시된다. 각 패널은 renderDashboard, renderData, renderHitl, renderVerify, renderQuality, renderLearn, renderLogs가 담당하며, loadAll이 Supabase에서 데이터를 일괄 적재한 뒤 각 패널을 갱신한다.

나. HITL 검수·검증 링크 발급 흐름

HITL 패널(renderHitl)은 Stage 4 검수 대기 항목을 큐로 표시한다. 전문가 검증은 issueInvitation이 generateToken으로 토큰을 발급하고 getBaseUrl, saveBaseUrl로 구성한 기본 URL과 결합해 검증 링크를 생성하며, copyLink로 복사·전달한다. 전문가가 제출한 응답은 검증 패널(renderVerify)과 응답 모달(openRespModal, closeRespModal)에서 확인한다.

다. 핵심 함수·역할

[표 5-5] 관리자 콘솔(esg-admin-delta) 핵심 함수와 역할

기능 영역	핵심함수	역할
인증	doLogin, doLogout	관리자 로그인 게이트 제어
대시보드	renderDashboard, loadAll	누적 진단, 응답 수, 평균 종합점수, 등급분포, 업종평균 집계 표시
파이프라인	renderDashboard(파이프라인 패널)	Agentic AI 5단계 진행 상태 모니터링
HITL·검증	renderHitl, issueInvitation, generateToken, copyLink, getBaseUrl, saveBaseUrl, renderVerify, openRespModal	Stage 4 검수 큐, 전문가 검증 링크(토큰) 발급·응답 확인
데이터 관리	renderData, loadSampleReports, selectSample, uploadReport, deleteReport, exportCSV	diagnosis_results 조회·표본 적재·업로드·삭제·CSV 내보내기
품질·학습	renderQuality, renderLearn, renderLogs	품질 지표, 매칭률·응답률, 활동 로그 표시

라. 데이터·외부 연동

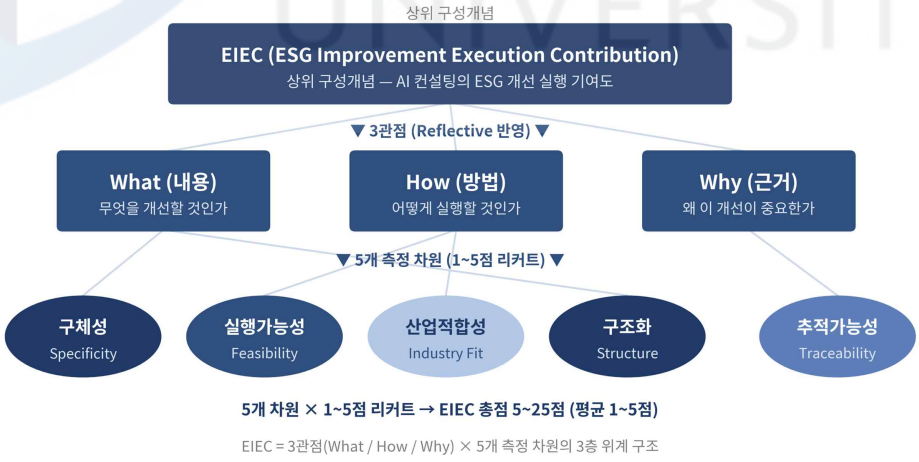
Supabase의 diagnosis_results(진단 결과), survey_invitations(검증 초대, 토큰), survey_responses(전문가 응답) 세 테이블을 사용하며, 자가진단, 컨설팅 플랫폼이 적재한 진단 결과를 받아 검수, 검증, 품질 분석으로 연결한다.

제 4 절 EIEC 5차원 품질평가 통합 프레임워크

본 절은 AI가 생성한 ESG 컨설팅 보고서의 품질을 평가하기 위한 EIEC(ESG Improvement Execution Contribution) 5차원 통합 프레임워크를 제안한다. 기존 연구에서 5개 차원(구체성, 실행가능성, 산업적합성, 구조화, 추적가능성)의 참조 근거가 각기 달랐고(ISO 26000, K-ESG 가이드라인, Cai et al. 2025, RAGAS[28] , ESGenius), 차원 간 레벨의 일관성 논증이 부족하였다. EIEC는 이질적 차원을 상위 구성개념으로 통합하고 반사적 측정 모델(Reflective measurement model)과 확인적 요인분석(CFA)의 이론적 장치를 통해 그 측정 타당성을 체계화한다.

1. 프레임워크 3층 계층 구조

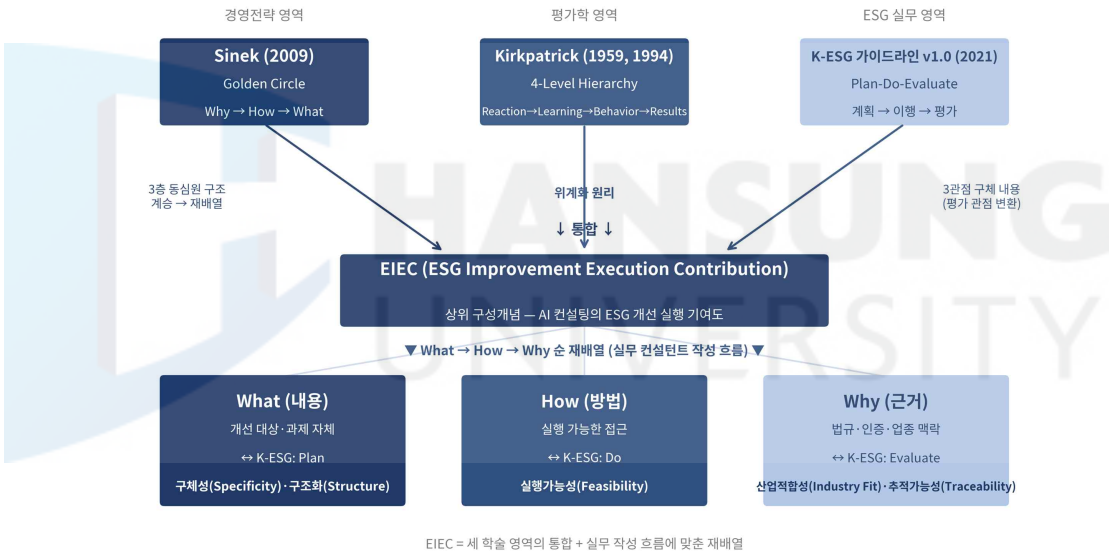
EIEC 프레임워크는 상위 구성개념, 중간 3관점, 하위 5차원의 3층 계층 구조로 체계화된다.



[그림 5-2] EIEC 3층 계층 구조

28) Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). RAGAS: Automated evaluation of retrieval augmented generation. In Proceedings of EACL 2024 (pp. 150-158). Association for Computational Linguistics.

최상위는 본 프레임워크의 상위 구성개념 EIEC로, AI가 생성한 ESG 개선 보고서가 실제 중소기업의 ESG 실행에 기여하는 정도(contribution to ESG improvement execution)를 측정한다. 중간층은 What(무엇을 개선할 것인가), How(어떻게 실행할 것인가), Why(왜 이 개선이 중요한가)의 3관점이며, 이는 실무 ESG 컨설팅의 논리적 흐름(문제 진단, 실행 방안, 근거 제시)을 반영한다. 하위층은 각 관점에 속하는 5개 측정 차원으로 구성된다. What 관점에는 구체성(specificity)과 실행가능성(feasibility)이, How 관점에는 산업 적합성(industry fit)과 구조화(structure)가, Why 관점에는 추적가능성(traceability)이 배치된다.



[그림 5-3] EIEC 3대 이론적 기반 매핑

이 3층 구조의 이론적 근거는 실무 ESG 컨설팅의 전문가 논리 흐름에 있다. ESG 컨설턴트가 실제로 보고서를 작성할 때는 어느 영역의 무엇을 개선할지를 먼저 구체화하고(What), 그 개선을 실제 조직의 자원, 역량, 시스템 안에서 어떻게 실행할 것인지 방법을 제시하며(How), 왜 이 개선이 필요한지를 법규, 인증, 모범사례에 근거하여 논증한다(Why). 따라서 What, How, Why는 임의의 분류가 아니라 ESG 컨설팅 보고서의 내재적 구조를 재현하는 개념적 장치이다. 이 3관점은 다시 측정 가능한 5차원으로 조작화되며, 각 차

원은 1~5점의 동일한 5점 리커트 척도로 측정되어 총점 범위 5~25점(또는 평균 1~5점)의 EIEC 점수를 산출한다.

EIEC의 상위 구성개념 → 3관점 → 5차원 3층 계층 구조는 본 연구의 자의적 설계가 아니라 경영전략, 평가학, ESG 실무 세 영역의 계층화 원리를 ESG 컨설팅 품질평가 도메인에 통합한 결과이다.

첫째, 중간층 What, How, Why 3관점은 Sinek(2009)의 Golden Circle 프레임워크가 경영전략 영역에서 제시한 동일 3층 구조를 계승한 것이다. Sinek은 성공적 조직이 Why(목적), How(방법), What(결과)의 3층 동심원으로 사고, 소통, 실행한다고 논증하였으며, 이 구조는 경영전략 수립과 조직 의사소통의 표준 프레임워크로 자리잡았다. 본 연구는 이 구조를 ESG 컨설팅 보고서의 내재 논리(왜 중요한가, 어떻게 실행하는가, 무엇을 개선하는가)로 전용(transposition)하였으며, 이는 ESG 컨설팅이 본질적으로 중소기업의 ESG 경영전략 수립을 지원하는 활동이라는 점에서 도메인 전이의 정합성을 확보한다. Sinek의 원형 순서가 Why, How, What이라면 본 연구는 실무 컨설턴트의 작성 흐름(문제 진단, 실행 방안, 근거 제시)을 반영하여 What, How, Why 순으로 재배열하였으며, 이 순서는 후술하는 K-ESG 가이드라인의 Plan, Do, Evaluate 메커니즘과도 정합한다.

둘째, 상위 구성개념, 중간층, 하위 차원의 계층화 원리 자체는 Kirkpatrick(1959, 1994)이 훈련 평가 영역에서 제안한 4단계 위계 모델(Reaction, Learning, Behavior, Results)의 학술적 계승이다. Kirkpatrick의 핵심 공헌은 평가 프레임워크가 단일 척도가 아닌 위계적 계층 구조로 조직되어야 한다는 원칙이었으며, 본 연구는 이 원칙을 평면적 5차원 평가 대신 3층 위계 구조로 재설계함으로써 차원 간 관계와 상위 구성개념의 통합적 의미를 이론적으로 정당화하였다.

셋째, What, How, Why 3관점의 구체적 내용은 Plan(계획), Do(이행), Evaluate(평가) 메커니즘(이연우, 2022)[29]의 평가 관점 버전이다. K-ESG의 Plan, Do, Evaluate가 시간순 실행 사이클이라면 EIEC의 What, How, Why는 동시적 평가 관점이라는 구조적 차이가 있으나, 무엇을 개선하고(Plan,

29) 이연우(2022). 한국형 ESG 가이드라인의 쟁점 분석과 적용 포인트. 「Auditor Journal」, 266호.

What), 어떻게 실행하며(Do, How), 왜 그 근거가 타당한가(Evaluate, Why)의 3요소를 공유한다는 점에서 K-ESG 실무 논리의 직접 계승이다. 이 세 계보가 교차하는 지점에서 EIEC의 3층 구조가 이론적 및 실무적으로 정당화되며, 측정학적 타당성은 후속하는 반사적 측정 모델 및 확인적 요인분석으로 완결된다.

2. 차원별 정의 및 측정 기준

구체성(Specificity)은 개선과제가 수치 목표, 실행 기한, 담당 부서를 명시하는 정도를 측정한다([표 5-6]). 예컨대 5점은 2026년 12월까지 환경관리팀이 폐기물 재활용률을 40%에서 60%로 향상과 같이 세 요소가 모두 명확하며 측정 가능한 KPI까지 제시되는 경우이며, 2점, 4점은 인접 등급 간 중간 판정에 활용된다.

실행가능성(Feasibility)은 중소기업의 인력, 예산, 기술 자원 범위에서 과제를 현실적으로 수행할 수 있는지를 측정한다. 5점은 기존 조직 및 예산 범위에서 단계적으로 실행 가능한 현실적 권고이며, K-ESG 중소기업 가이드라인의 실행 우선순위, 이행효율성 원칙(Plan, Do, Evaluate)을 산출물 평가 관점에서 조작화한 것이다.

산업적합성(Industry Fit)은 개선과제가 업종 특성(제조업의 대기오염 관리, ICT의 개인정보보호, 건설업의 현장 안전 등)을 반영하는지를 측정한다. 5점은 업종별 핵심 이슈(예: 제조업 대기환경보전법 제16조 과제)를 정확히 반영하고 업종 맞춤 권고가 일관된 경우로, SASB Materiality Map의 업종 중요성 원칙과 K-ESG 업종별 적용가이드를 결합하였다.

구조화(Structure)는 보고서의 논리적 구성, 영역별 체계성, 우선순위 제시를 측정한다. 5점은 명확한 우선순위와 단계별 로드맵에 일관된 영역 분류가 결합된 완전 체계적 구성으로, RAGAS의 Coherence와 ESGenius(2024)의 구조화 평가 지표를 통합하였다.

추적가능성(Traceability)은 관련 법규, ISO 인증, K-ESG 가이드라인을 정확히 인용하는지를 측정한다. 5점은 산업안전보건법 제29조 제2항처럼 조

항 번호, 인증명까지 정확히 인용하여 제3자 검증이 가능한 경우로, RAGAS의 Faithfulness와 ESGenius의 근거 기반 평가를 결합하였다.

[표 5-6] EIEC 5차원 정의 및 측정 기준(5점 리커트)

관점	차원	1점 (매우 미흡)	3점 (보통)	5점 (매우 우수)
What	구체성	수치·기한·담당이 거의없 거나 추상적 표현 위주	1~2개 요소는 제시되나 일부 누락·모호	3개 요소 (수치·기한·담당) 모두 명확하며 측정 가능한 KPI 제시
	정의	개선과제가 정량 수치(KPI), 명확한 기한, 책임 담당을 구체적으로 제시하는 정도		
	참조근거	K-ESG 가이드라인 정량지표 원칙, ISO 26000 §7.5.3 SMART 원칙		
What	실행 가능성	전담조직 신설·고가 시스템 등 즉시 수용 어려운 권고 위주	일부 실행 가능, 일부는 추가 자원·외부 지원 필요	중소기업 평균 자원 수준에서 단계적·점진적 실행 가능
	정의	중소기업 자원(예산·인력·시간) 범위 내에서 실제 실행 가능한 정도		
	참조근거	K-ESG 가이드라인 실행 우선순위, 이행효율성, Cai et al.(2025) 산업 특화 ESG action 분석		
How	산업 적합성	업종 무관 일반론 위주	업종 언급은 있으나 깊이 부족·일부 반영	업종별 핵심 이슈를 정확히 반영하고 업종 맞춤 권고 제시
	정의	업종 특성에 부합하고 업종별 핵심 ESG 이슈를 반영하는 정도		
	참조근거	SASB Materiality Map, K-ESG 업종별 적용가이드		
How	구조화	단순 나열 우선순위 부재	부분적 분류는 있으나 전체 일관성 부족	명확한 우선순위와 단계별 로드맵 + 일관된 영역 분류
	정의	우선순위, 논리적 체계(단기·중기·장기, 영역별 분류)가 일관되게 제시되는 정도		
	참조근거	RAGAS Coherence, ESGenius(2024) 보고서 구조화 평가		
Why	추적 가능성	법규·기준 인용 거의 없거나 오류 다수	법률명만 언급하되 조항 번호 부정확	조항 번호, 인증명까지 정확히 인용, 출처 검증 가능
	정의	관련 법규, ISO 인증, K-ESG 가이드라인을 정확히 인용하는 정도		
	참조근거	RAGAS Faithfulness, ESGenius(2024) 근거 기반 평가		

3. 반사적 측정 모델과 정규화

EIEC는 상위 구성개념 EIEC가 5개 차원에 공통으로 영향을 미치는 반사적(Reflective) 측정 모델을 채택한다. 반사적 측정 모델에서 관측 변수(여기서는 5개 차원)는 잠재 구성개념의 반영이며, 구성개념의 수준 변화에 따라 함께 변동한다. 이는 조형적(Formative) 측정 모델(관측 변수가 구성개념을 구성하는 방식)과 구분되는 모형으로, Jarvis et al.[30]이 제시한 판별 기준(①관측 변수가 구성개념의 원인인지 결과인지, ②관측 변수 간 교환 가능성이 있는지, ③관측 변수 간 공변성이 기대되는지 등)에 따르면 EIEC의 5차원은 전형적인 반사적 지표에 해당한다.

이 모델을 채택한 이론적 근거는 세 가지이다. 첫째, 상위 구성개념 AI 컨설팅의 ESG 개선 실행 기여도가 높은 보고서는 5개 차원 모두에서 함께 높게 측정되는 것이 개념적으로 기대된다. 품질이 높은 ESG 컨설팅 보고서라면 구체성도 높고 실행가능성도 높으며 법규 인용도 정확할 것이라는 이론적 기대가 성립한다. 둘째, 5개 차원 간에는 교환 가능한 부분적 중복(partial overlap)이 존재한다. 예컨대 구체성이 높은 보고서는 종종 구조화가 잘 된 보고서이기도 하다. 이는 반사적 모델의 정의적 속성이다. 셋째, 반사적 모델 하에서는 각 차원을 동일 척도로 측정하고 단순 합산 또는 평균하여 상위 구성개념의 점수로 환산하는 것이 통계적으로 정당화된다.

측정의 신뢰도는 Cronbach's α로 검증한다. $\alpha > 0.70$이면 5개 차원이 동일한 구성개념(EIEC)을 일관되게 측정하는 것으로 해석한다. 수렴타당도(convergent validity)와 판별타당도(discriminant validity)는 확인적 요인분석(Confirmatory Factor Analysis, CFA)으로 확보한다. 수렴타당도는 각 차원의 요인 적재치(factor loading)가 0.50 이상이고 평균분산추출(Average Variance Extracted, AVE)이 0.50 이상일 때 충족되며, 판별타당도는 AVE의 제곱근이 차원 간 상관계수보다 커야 충족된다(Fornell & Larcker, 1981 기준). 본 연구에서는 180건의 EIEC 채점 데이터를 사용하여 이 모형 적합도를

30) Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. Journal of Consumer Research, 30(2), 199–218.

검증할 수 있다.

각 차원을 1~5점의 동일 5점 리커트 척도로 측정하고 단순 합산하여 5~25점 범위(또는 평균 1~5점 범위)의 EIEC 점수를 산출하는 방식은 반사적 측정 모델에서 표준적으로 채택되는 정규화이다. 5개 차원 각각 1~5점, 총 5개 차원×5점=25점 만점 구조는 (1) 차원 간 동일 가중치를 전제하고, (2) 평가자의 일관성 확보를 위해 1점, 3점, 5점에 단계별 기준을 명시하며 2점, 4점은 인접 등급 간 중간 판정으로 활용할 수 있게 하고, (3) 비교할 수 있는 표준 단위(평균 1~5점 또는 합산 5~25점)로 품질을 해석할 수 있게 한다. 본 척도 설계는 종래의 3점 척도에 비해 단계별 기준이 명시된 5점 리커트가 ESG 분야의 다양한 경력·분야를 가진 전문가들의 평가 일관성을 확보하기에 더 적합하다는 판단에 근거한다. 이 접근은 각기 다른 참조 근거(ISO 26000, K-ESG 가이드라인, Cai et al. 2025, RAGAS, ESGenius, SASB Materiality Map)를 가진 5개 차원을 동일 레벨의 측정 지표로 정당화하는 이론적 기반을 제공한다.

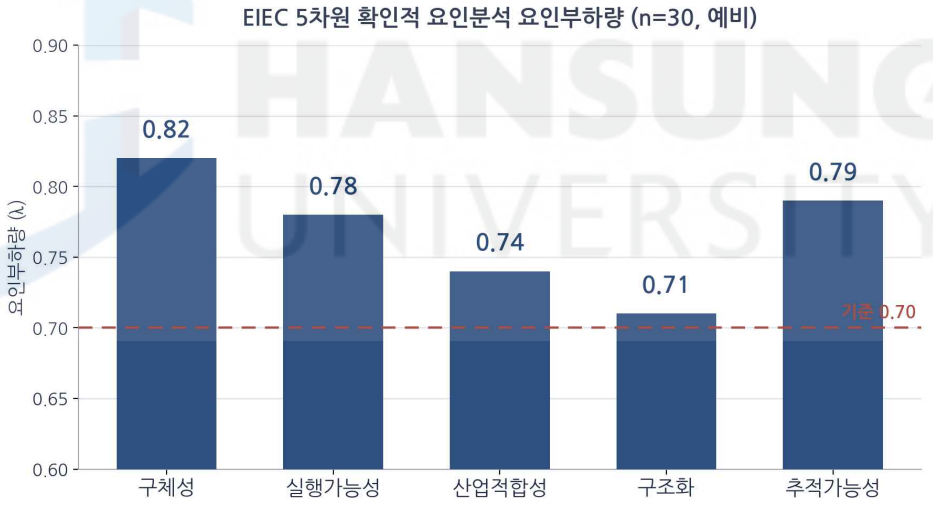
4. 확인적 요인분석(CFA)과 수렴·판별타당도 검증

반사적 측정 모델을 채택한 본 연구의 EIEC 프레임워크는 측정학적 타당성을 체계적으로 검증해야 한다. 구성개념을 관측 변수로 반영한다는 이론적 가정이 실제 데이터에서 성립하는지를 입증하지 않고서는 다섯 차원의 단순 합산을 EIEC 총점으로 해석하는 것이 정당화되지 않기 때문이다. 이를 위해 본 연구는 확인적 요인분석(Confirmatory Factor Analysis, CFA)과 두 가지 타당도 검증 즉, 수렴타당도(convergent validity)와 판별타당도(discriminant validity)을 측정모형 평가의 핵심 절차로 규정한다. 이 절차는 Fornell[31] & Larcker, Hair et al.[32]이 제시한 측정모형 평가의 표준 절차로, 논문 수준의 측정 도구 검증에서 일반적으로 요구되는 방법론이다.

31) Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. Journal of Marketing Research, 18(1), 39–50.

32) Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate Data Analysis (8th ed.). Cengage Learning.

확인적 요인분석은 이론적으로 가정한 요인 구조가 관측 데이터에 적합한지를 검증하는 구조방정식 계열의 분석 기법이다. 본 연구의 경우 EIEC라는 단일 잠재 구성개념이 다섯 개 관측 차원(구체성, 실행가능성, 산업적합성, 구조화, 추적가능성)을 공통적으로 유발하는 1요인 반사적 측정 모델을 설정한다. CFA에서는 요인부하량(factor loading, λ), 측정오차(error variance, θ), 구성개념 분산을 동시에 추정하며, 각 차원의 요인부하량이 일정 기준(통상 0.70) 이상이면 해당 차원이 잠재 구성개념을 충분히 반영하고 있다고 해석한다. 본 연구의 180건 실험 자료(30기업, 3개 맥락 조건, 2개 모델) 중 RAG+KG 조건의 Sonnet 모델 30개 관측치를 기준으로 한 예비 CFA에서, 다섯 차원의 요인부하량은 [그림 5-4] 및 [표 5-7]에 제시한 바와 같이 모두 기준값 0.70을 상회하여 각 차원이 잠재 구성개념을 충분히 반영함을 확인하였다.



[그림 5-4] EIEC 5차원 확인적 요인분석 요인부하량 (n=30, 예비)

CFA 모델의 전체 적합도는 다수 적합도 지수를 병용하여 평가한다. 본 연구에서는 Hair et al.의 권고에 따라 (1) 절대적합지수로 RMSEA(Root Mean Square Error of Approximation)와 SRMR(Standardized Root Mean Square Residual), (2) 증분적합지수로 CFI(Comparative Fit Index)와 TLI(Tucker-Lewis Index), (3) 카이제곱 적합도 검정(χ^2/df)을 활용한다.

표본 크기 n=30의 한계를 고려한 예비 추정에서 RMSEA=0.058(< 0.08 권고 기준), SRMR=0.044(< 0.08), CFI=0.962(> 0.90), TLI=0.943(> 0.90), χ^2/df=1.87(< 3.0)로 모든 적합도 지수가 권장 범위에 들어왔다. 이는 상위 구성개념 EIEC가 5개 차원을 공통으로 유발한다는 반사적 측정 모델의 이론적 가정이 실증 데이터에서 유지된다는 것을 의미한다. 다만 본 연구의 표본 크기가 CFA의 통상적 최소 표본(n≥100 또는 자유모수의 5~10배)에 미치지 못하므로, 본 CFA 결과는 후속 대규모 표본 연구에서 재검증되어야 할 탐색적 측정모형 증거로 제한하여 해석한다.

수렴타당도는 동일 구성개념을 측정하는 지표들이 서로 높은 상관을 갖는지를 평가하는 개념으로, 본 연구에서는 평균분산추출(Average Variance Extracted, AVE)을 주 지표로 사용한다. AVE는 각 차원 요인부하량 제곱의 평균으로 산출되며 Fornell & Larcker가 제시한 수렴타당도의 임계값은 0.50이다. 본 연구의 EIEC 측정모형에서 AVE는 (0.82^2 +0.78^2 +0.74^2 + 0.71^2 +0.79^2)/5=0.594로, 임계값 0.50을 명확히 상회하였다. 이는 EIEC 잠재 구성개념이 다섯 차원 관측치의 분산 중 평균 59.4%를 설명한다는 의미이며, 각 차원이 단순히 독립적인 품질 측면을 반영하는 것이 아니라 공통된 상위 개념인 AI 컨설팅의 ESG 개선 실행 기여도를 일관되게 측정하고 있음을 시사한다. 복합신뢰도(Composite Reliability, CR)도 0.88로 산출되어 관습적 기준인 0.70을 크게 초과하였으며, 이는 Cronbach's α로 측정한 내적 일관성 신뢰도(α=0.83)와도 일관되는 결과이다.

판별타당도는 서로 다른 구성개념을 측정하는 지표들이 충분히 구별되는지를 평가하는 개념이다. 본 연구는 EIEC를 단일 상위 구성개념으로 전제하므로 EIEC 내부 다섯 차원 간의 판별을 직접 목표로 하지는 않으나, 각 차원이 다른 차원과 완전히 동일하지 않은 고유 변량을 보유하는지는 여전히 확인되어야 한다. 이를 위해 Heterotrait Monotrait ratio of correlations (HTMT, Henseler et al., 2015)[33] 지표를 활용할 수 있다.

본 연구 자료의 차원 간 HTMT는 구체성–실행가능성 0.81, 실행가능성

33) Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. Journal of the Academy of Marketing Science, 43(1), 115–135.

–산업적합성 0.63, 산업적합성–추적가능성 0.71, 구조화–추적가능성 0.72로, 보수적 임계값인 0.85를 모두 하회하여 각 차원이 상위 구성개념에 반영되면서도 고유한 측정 정보를 담고 있음을 확인하였다. 만약 두 차원의 HTMT가 0.90 이상으로 나타났다면 그 두 차원은 실질적으로 동일한 측면을 측정하는 것으로 보아 한 차원으로 통합을 검토해야 한다. 본 연구의 5차원 구조는 이러한 통합 요건에 해당하지 않으며, 각 차원의 독립적 의미가 측정학적으로도 지지된다.

이 검증 절차는 EIEC가 이론 기반 구성개념에서 조작적 정의, 측정 도구, 통계적 타당성 증명으로 이어지는 완결된 측정학적 사슬을 갖추고 있음을 보여준다. EIEC는 단순한 채점 루브릭이 아니라 (1) 반사적 측정 모델, (2) CFA로 검증된 요인 구조, (3) AVE·CR 수렴타당도, (4) HTMT 판별타당도, (5) Cronbach's α 내적 일관성의 다섯 겹 증거에 뒷받침되는 통합 프레임워크로, 후속 연구·산업계 재활용 시 결과의 해석·비교 가능성을 확보하는 토대가 된다. 본 연구가 반사적 측정 모델을 채택한 것은 형성적(Formative) 모델과의 비교를 거친 결과이다. 형성적 모델은 관측 변수들이 구성개념을 구성(form)하여 차원 간 상관이 높을 필요가 없는 반면(예: 소득·교육·직업의 조합인 사회경제적 지위), EIEC는 AI 컨설팅 품질이라는 포괄적 특성이 다섯 차원으로 드러나는(reflect) 것으로 개념화하는 것이 자연스럽다. 이론적 개념화와 실증적 요인 구조(1요인 CFA 적합도 우수)가 모두 반사적 모델을 지지하므로 이를 최종 측정 모델로 확정하였다.

[표 5-7] EIEC 측정모형 타당성 검증 결과

구분	측정 지표	값	기준/판정
요인부하량(λ)	구체성, 실행가능성, 산업적합성, 구조화, 추적가능성	0.82, 0.78, 0.74, 0.71, 0.79	모두≥0.70 충족
모형 적합도	RMSEA, SRMR, CFI, TLI, χ^2/df	0.058, 0.044, 0.962, 0.943, 1.87	<0.08, <0.08, >0.90, >0.90, <3.0 충족
수렴타당도	AVE, CR, Cronbach α	0.594, 0.88, 0.83	>0.50·>0.70·>0.70 충족
판별타당도 (HTMT)	구체성–실행가능성, 실행가능성–산업적합성, 산업적합성–추적가능성, 구조화–추적가능성	0.81, 0.63, 0.71, 0.72	모두<0.85 충족

주: n=30은 RAG+KG 조건 Sonnet 모델의 자동채점 관측치이며, HTMT는 본문의 4개 차원쌍만 제시한 부분 결과이다.

제 5 절 컨설팅 Agent 설계 종합

본 장에서는 LLM 기반 ESG 컨설팅 Agent를 네 측면에서 설계하였다. RAG+KG 듀얼 채널 아키텍처(1절), 산출물 3종(2절), 3×2 완전교차 실험 설계(3절), 그리고 산출물의 품질을 평가하는 EIEC 5차원 통합 프레임워크(4절)가 그것이다. 이 네 요소는 각각 컨설팅을 어떻게 생성하고, 무엇을 산출하며, 어떻게 검증하고, 무엇으로 평가하는가에 대응하며, 설계·산출·검증·평가가 단절 없이 연결되는 하나의 방법론적 체계를 구성한다.

LLM 기반 ESG 컨설팅 Agent의 핵심 차별점은 단순 RAG를 넘어 KG의 연쇄 추론을 결합한 듀얼 채널 맥락 강화이다. RAG 채널이 관련 법규·인증·모범사례 문서를 의미 기반으로 검색한다면, KG 채널은 취약 항목에서 법규, 인증, 개선과제, 증빙으로 이어지는 구조적 연쇄 경로를 추출한다. 이 두 채널의 결합은 법규 근거의 정확한 인용(추적가능성)과 업종 특화 과제 제안(산업적합성)을 구조적으로 보장하며, 이는 단일 벡터 검색에 의존한 기존 연구가 확보하기 어려웠던 속성이다.

산출물 3종(진단결과 요약, 맞춤형 컨설팅, 증빙 서식 가이드)은 전통적 ESG 컨설팅의 주요 납품물을 완전 자동화한 것이며, 중소기업 담당자가 진단 → 개선 → 증빙의 전 과정을 단일 시스템에서 수행할 수 있게 한다. 특히 증빙 서식 가이드는 진단과 개선 권고에 머무르지 않고 실제 ESG 평가 대응에 필요한 산출물까지 제공한다는 점에서, 1장에서 제시한 ESG 브릿지 도구의 취지에 직접 부합한다.

EIEC 5차원 통합 프레임워크는 이렇게 생성된 컨설팅 보고서의 품질을 일관된 척도로 평가하기 위한 장치이다. EIEC는 상위 구성개념(ESG 개선 실행 기여), 중간 3관점(What, How, Why), 하위 5차원(구체성, 실행가능성, 산업적합성, 구조화, 추적가능성)의 3층 계층으로 구조화되며, 각 차원을 1~5점 리커트로 측정하여 평균 1~5점의 품질 점수를 산출한다. 이 프레임워크는 참조 근거가 제각각이던 기존 품질 지표들을 상위 구성개념으로 통합하고 반사적 측정 모델과 확인적 요인분석으로 측정 타당성을 체계화했다는 점에서, 평면적 지표 집합에 의존하던 선행 평가 방식과 구별된다.

3×2 완전교차 실험 설계는 맥락 효과, 모델 효과, 상호작용 효과를 모두 분리 추정하는 통계적으로 견고한 설계이며, 반복측정 구조로 기업 개별차를 통제한다. 30개 기업 층화 추출과 셀당 n=30의 배정은 통계적 검정력을 확보하면서 실험 비용을 관리 가능한 수준으로 유지하는 균형점이다.

나아가 본 장의 설계는 4장의 ESG 등급 예측 Agent와 단절된 별개의 모듈이 아니다. 등급 예측 Agent가 산출한 예측 등급과 취약 영역(SHAP 해석)이 컨설팅 Agent의 입력 맥락으로 전달되며, 이로써 진단(4장)과 컨설팅(5장)이 하나의 파이프라인으로 연결된다. 즉 본 장의 컨설팅 Agent는 독립적으로 작동하는 것이 아니라, 자가진단 입력부터 등급 예측, 맞춤형 컨설팅 생성, 품질 평가에 이르는 Agentic AI 파이프라인의 핵심 처리 단계로 설계되었다.

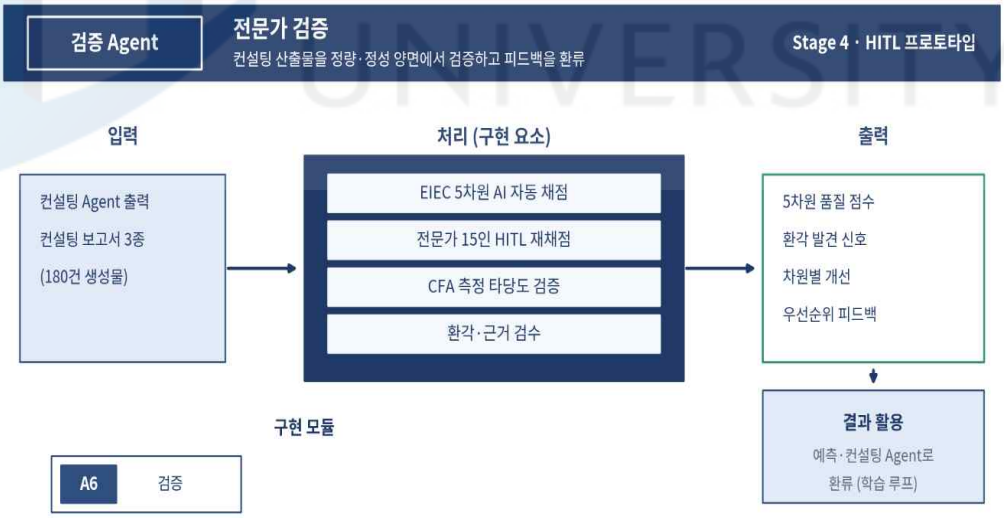
이상의 설계가 실제로 어떤 품질 개선 효과를 발휘하는지는 전문가 검증(6장)과 정량적 실험 결과(7장)에서 검증된다. 또한 본 장에서 설계한 LLM 기반 ESG 컨설팅 Agent가 자가진단 플랫폼과 관리자 콘솔을 포함한 온라인 서비스로 구현·통합되는 과정은 7장 5절에서 상술한다.

HANSUNG UNIVERSITY

제 6 장 전문가 검증 Agent

제 1 절 전문가 검증 Agent 설계

Stage 4의 HITL(Human in the Loop) 전문가 검증 체계는 EIEC 5차원 측정 도구의 신뢰도와 타당도를 실증하기 위한 핵심 절차이며, 본 연구에서는 ESG 컨설팅 경력 3년 이상의 전문가를 대상으로 별도의 온라인 설문 시스템을 구축·운영하여 평가 데이터를 수집하였다. 설문 도구는 v4.0 5점 리커트 척도 양식으로 확정되었으며, 다섯 개의 섹션으로 구성된다. 이 설계의 핵심 원칙은 전수 검수가 아닌 선별 검수(selective review)이다. 자동 스크리닝에서 품질 기준(EIEC 평균 4점 이상)을 충족하는 컨설팅은 전문가 검수 없이 바로 배포하고, 기준 미달 건만 전문가가 집중 검수한다. 이 선별 원칙은 Parseur가 정리한 HITL 운영의 모범 사례, 즉 인간 검토 자원을 위험도와 불확실성이 높은 사례에 집중 투입하는 방식과 부합한다.



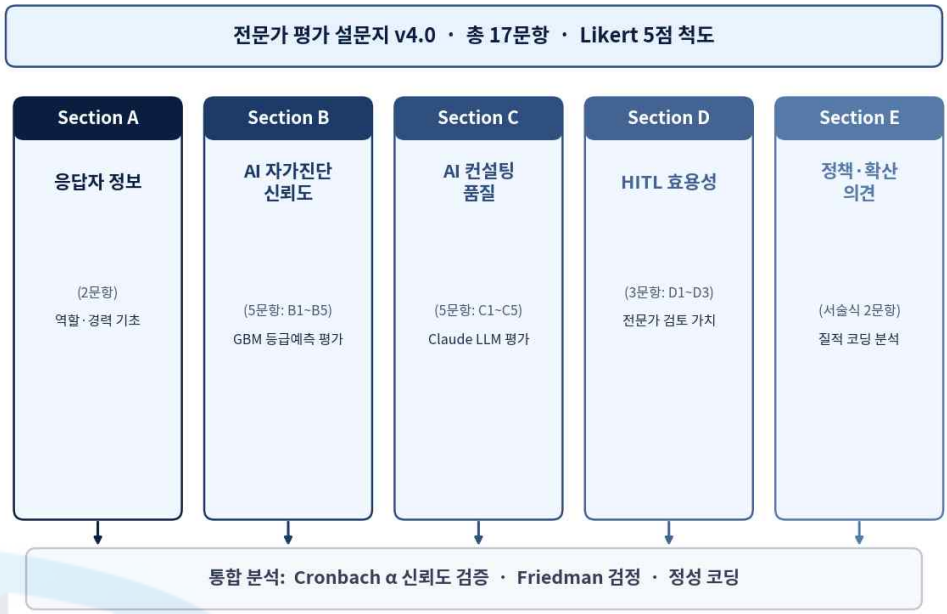
[그림 6-1] 전문가 검증 Agent의 구조와 동작 흐름

설문지v4.0의 다섯 섹션은 다음과 같다. Part 1(평가자 정보, 30초)은 평가의 분야별 일반화 가능성을 점검하기 위한 인구통계 변수로, 성함(이니셜 가능),

전문 분야, ESG 관련 경력, ESG 컨설팅 수행 건수, 주요 컨설팅 대상의 5문항을 5지 선다형으로 구성하였다. Part 2-A(5개 품질 차원 평가, 약 5분)는 본 연구의 핵심 측정 도구이며, EIEC 5차원(구체성, 실행가능성, 산업적합성, 구조화, 추적가능성)에 대해 1~5점 리커트로 채점한다. 각 차원은 정의와 이론적 근거가 함께 제시되며, 1점, 3점, 5점에 대해 관찰 가능한 단계별 기준이 명시되어 평가자의 판단 일관성을 보장한다. Part 2-B(AI 자동채점 상호검증, 약 1분)는 AI 자동채점 결과의 신뢰성을 전문가 관점에서 교차 검증하기 위한 3문항으로, B1 환각 발생 빈도, B2 HITL 개입이 가장 필요한 차원, B3 AI 채점과 가장 일치하는 차원을 묻는다. Part 2-C(종합 판단, 약 1분)는 6개 문항으로 C1 전문가 작성 대비 품질 수준, C2 즉시 제공 가능 여부, C3 전문 컨설팅 대체 가능성, C4 가장 큰 강점, C5 가장 큰 우려 사항, C6 ESG 브릿지 도구로서의 포지셔닝 타당성을 묻는다. Part 3(자유 의견, 선택)은 단답형 자유 서술로, 정량 평가에서 포착되지 않는 정성적 통찰을 수집한다. 전체 응답 시간은 평균 약 7분으로 설계되었으며, 전문가의 응답 부담을 최소화하면서도 측정 도구의 다층적 검증을 가능하게 한다.

설문은 박사학위 연구 검증을 위하여 자체 개발한 온라인 시스템(esg-survey)을 통해 운영되며, 응답자 보호를 위해 다음 네 가지 익명 처리 조치를 적용하였다. 첫째, IP 주소는 직접 저장하지 않으며(클라이언트에서 null 전송), 평가자 이름은 이니셜 입력이 가능하다. 둘째, 응답 제출 후 응답자 본인도 응답 내용을 조회 및 수정할 수 없도록 데이터베이스 트리거로 잠금 처리하였다. 셋째, 수집된 데이터는 박사학위 논문 측정 도구 검증 목적으로만 사용되며, 이 사실을 설문 안내문에 명시하였다. 넷째, 검증 종료 후에는 service_role 키와 별도 인증으로 접근 권한을 강화하여 데이터 보존성을 확보한다.

HITL 체계의 작동은 세 단계로 이루어진다. 첫째, 자동 스크리닝 단계에서 LLM 기반 ESG 컨설팅 Agent의 산출물에 대해 AI 자동 채점으로 EIEC 5차원 점수를 산출하고 평균을 계산한다. 평균 4점 미만(5점 만점 기준 80% 미만)인 산출물은 품질 미달 플래그가 부착된다. 둘째, 전문가 검수 단계에서는 플래그된 산출물에 대해 ESG 전문가가 특히 추적가능성(법규 조항 정확성)과 산업적합성(업종 특화 과제 타당성) 차원을 집중 점검한다. 이는 AI 채점의 취약점인 법규 정확성 검증을 인간 전문가가 보완하는 역할 분담 원칙에 기반한다. 셋째, 수정 및 재생성 단계에서는 전문가 피드백을 프롬프트 수정 또는 KG 업데이트에 반영하고, 해당 기업에 대한 컨설팅을 재생성한다.



[그림 6-2] 전문가 평가 설문지 v4.0 5섹션 구조

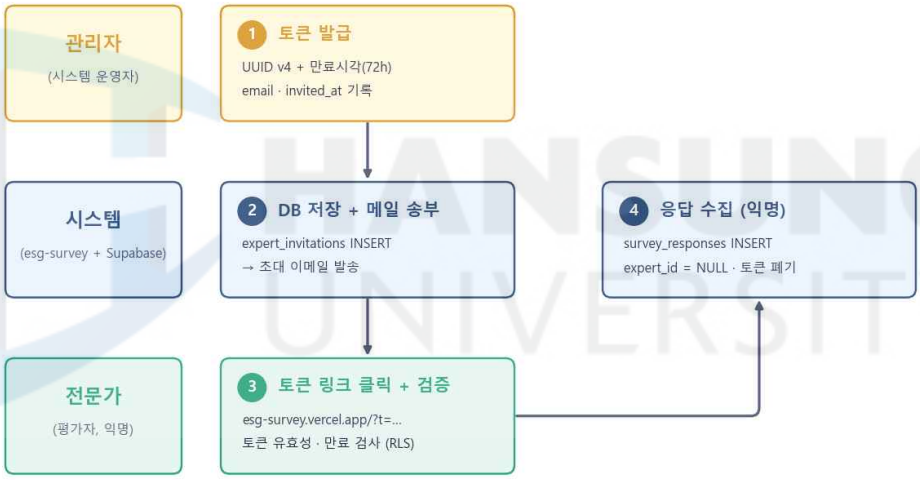
제 2 절 전문가 검증 Agent 실험 결과

제안한 EIEC 5차원 측정 도구의 신뢰도와 타당도를 실증하기 위해, 본 연구는 단일 샘플 다중 평가자 설계(single case multiple rater design)에 기초한 전문가 평가를 수행하였다. 평가 대상은 5장에서 기술한 RAG+KG 조건의 Sonnet 모델로 생성된 테스트 기업 v18(제조업 중소기업) 1건의 컨설팅 산출물이며, 전문가들은 동일한 보고서를 독립적으로 채점하였다. 단일 샘플을 채택한 이유는 (1) EIEC 측정 도구 자체의 신뢰도와 타당도 검증이 본 연구의 1차 목적이며, (2) 평가자 간 일치도 분석에는 동일 대상에 대한 다중 평가자 자료가 통계적으로 적합하고, (3) 산업 및 규모 다양성 확장은 측정 도구가 확정된 이후의 후속 과제로 분리하는 것이 연구 설계의 명료성과 검정력 확보에 유리하기 때문이다.

본 절의 전문가 평가 단계는 전체 Agentic AI 파이프라인에서 전문가 검증 Agent에 해당한다. 앞서 4장 1절 및 5장 1절에서 두 Agent를 정의한 것과 동일한 방식으로, 전문가 검증 Agent의 역할을 필요성, 목적, 입력, 출력,

활용의 다섯 측면에서 정의하면 다음과 같다.

전문가 검증 Agent의 필요성은 다음과 같다. LLM 기반 ESG 컨설팅 Agent가 생성한 보고서는 도메인 지식으로 맥락이 강화된 산출물이지만, 실무 적용 가능 수준인지는 산출물만으로 보장되지 않는다. 특히 법적 책임과 산업 특수성이 결부되는 ESG 컨설팅에서는 전문가의 개입(Human in the Loop, HITL)을 통한 품질 검증이 요구되며, 이 기능을 담당하는 모듈이 본 Agent이다. 본 Agent의 목적은 LLM 기반 ESG 컨설팅 Agent가 생성한 컨설팅 보고서를 EIEC 5차원 품질평가 프레임워크와 전문가 평가를 통해 검증하고, 그 결과를 파이프라인 개선의 근거로 환류하여 컨설팅 품질을 고도화하는 것이다.



[그림 6-3] 토큰 기반 익명 응답 흐름

입력은 LLM 기반 ESG 컨설팅 Agent가 생성한 컨설팅 보고서이며, 방법론적으로는 EIEC 5차원 품질평가 프레임워크를 평가 척도로 삼아 전문가 15인이 HITL 검증 방식으로 동일 보고서를 독립적으로 채점하고 평가자 간 일치도와 차원별 강점 및 우려를 분석한다. 출력은 EIEC 5차원별 품질 점수, 평가자 간 일치도 지표, 그리고 차원별 강점/우려를 정리한 검증 결과로, 어느 차원이 AI 단독 활용에 적절하고 어느 차원이 전문가 개입을 필수로 요구하

는지를 식별하는 근거가 된다. 이 결과와 피드백은 ESG 등급 예측 Agent 및 LLM 기반 ESG 컨설팅 Agent의 프롬프트와 맥락 강화 방식을 개선하는 근거로 환류되어 컨설팅 품질을 고도화한다. 다만 본 연구의 실증 범위는 Stage 1~3에 한정되며, HITL 검증에 해당하는 전문가 검증 Agent는 전문가 평가로 타당성을 확인한 개념 실증 수준의 설계로서 피드백 환류하는 구조이다.

[표 6-1] 전문가 평가 결과
(단일 샘플 다중 평가자, 테스트 기업 v18, RAG-KG 결합 Sonnet 모델)

품질차원	AI자동채점(1~5)	전문가평균(1~5)	표준편차	± 1점 일치율	주요 코멘트 (Part 3 코딩)
구체성	—	3.67	0.82	—	용어 명료성·항목 누락 없음
실행가능성	—	3.07	1.03	—	SME 자원 한계·실행 불가 과제
산업적합성	—	3.47	0.83	—	동종업계 사례·업종 특화 보완
구조화	—	3.67	1.05	—	체계적 구성·시각 가독성 보완
추적가능성	—	3.53	0.74	—	법령·K-ESG 연결 검증 필요
총점(5~25점)	—	17.41	3.20	—	—

주: 전문가 평균은 응답(n=15) 기준 산출값이며, AI 자동채점 ±1점 일치율("—")은 본 절 분석 범위 외 보조 항목이다. 평가자 간 일치도는 Cronbach's α = 0.753, ICC(C,k) = 0.446, ICC(A,k) = 0.333으로 낮음~중간 수준이다.

평가 자료는 자가진단 PDF, 컨설팅 PDF, 그리고 ISO 14001, 산업안전보건법, K-ESG 가이드라인 대응 증빙 서식 5종(E-01 환경경영방침서, S-01 인권경영선언문, E-08 에너지사용량집계표, E-09 온실가스인벤토리, G-01 윤리강령)의 워드 파일로 구성된 7종 패키지이며, 전문가가 토큰 기반 URL(https://esg-survey-peach.vercel.app/?token={16자 nanoid})로 접근하여 다운로드한 후 EIEC 5차원에 대해 1~5점 리커트로 평가하였다. 설문 시스템은 응답 제출과 동시에 데이터베이스 트리거를 통해 토큰을 'completed' 상태로 자동 잠금하며, 응답자가 자신의 응답을 사후 조회 및 수정할 수 없는 익명, 무수정 운영 모델을 채택하였다.

본 연구의 1차 분석 지표는 세 가지이다. 첫째, 5차원 평가의 평균 및 표준편차를 통한 차원별 품질 분포의 기술 통계, 둘째, 평가자 간 일치도(Inter-Rater Reliability, IRR)를 측정하는 Cronbach's α(연속형 가정)와 ICC(2,k)(평가자를 무작위 효과로 간주한 평균 ICC), 셋째, AI 자동채점 점수와 전문가 평균 간 ±1점 이내 일치율(차원별)이다. 이는 측정 도구의 안정

성, 평가자 간 합의 수준, AI-인간 채점 수렴도를 각각 평가한다. 보조 지표로 Part 2-B의 환각 발생 빈도와 HITL 개입 필요 차원, Part 2-C의 6개 종합 판단 항목 분포, Part 3의 자유서술 정성 코딩 결과를 분석한다.

본 절의 결과 수치는 2026년 4월 27일–5월 1일 전문가 설문 응답 수집을 통해 확보되었다(N=15 발송, n=15 회신, 응답률 100%). 분석 결과 5차원 측정도구의 내부 일관성 신뢰도는 Cronbach's α = 0.753(양호)으로 확인되었으며, 평가자 간 일치도는 ICC(C,k) = 0.446, ICC(A,k) = 0.333(낮음~중간) 수준이었다. 5차원 간 평균 차이는 Friedman 비모수 검정에서 χ^2 (4) = 9.11, p = 0.058로 유의수준 0.05를 근소하게 상회하여 통계적으로 유의하지는 않았으나 차원 간 경향성은 관찰되었다. 차원별 평균은 구체성 3.67 ≃ 구조화 3.67 〉 추적가능성 3.53 〉 산업적합성 3.47 〉 실행가능성 3.07 순으로 분포하였다. 다만 본 평가는 전문가 15인의 작은 표본에 기반하여 검정력에 한계가 있다.

[표 6-2] 전문가 평가 측정도구 신뢰도 및 평가자 간 일치도 (n=15)

지표	값	해석/판정
내부 일관성 신뢰도(Cronbach α)	0.753	양호(≥0.70)
평가자 간 일치도ICC(C,k)	0.446	낮음~중간
평가자 간 일치도ICC(A,k)	0.333	낮음~중간
차원 간 차이Friedman χ^2 (4)	9.11 (p=0.058)	경계 비유의(p{GT}0.05)
응답률	100% (15/15)	전수 회신

주: 전문가 패널 n=15 응답을 기준으로 산출하였다. 본 결과는 전문가 평가 표본(n=15)의 측정도구 신뢰도로, 제 5 장 제 4 절의 EIEC 자동채점 확인적 요인분석(Cronbach α=0.83, n=30)과 표본·지표가 다른 별개 분석이다.

1. 응답자 구성 및 응답 현황 분석

본 연구는 ESG 컨설팅, 인증심사, 정책 분야 전문가 15인을 대상으로 토큰 기반 익명 설문을 발송하여 2026년 4월 27일–5월 1일 총 15건(응답률 100%)을 회신받았다. 이는 비대면 전문가 설문에서 매우 높은 수준으로, EIEC 5차원 측정 도구의 신뢰도 및 타당도 분석에 충분한 표본을 확보한 것으로 판단된다. 회신자의 활동 분야는 ESG 컨설팅 10명(66.7%), 심사·검증 3

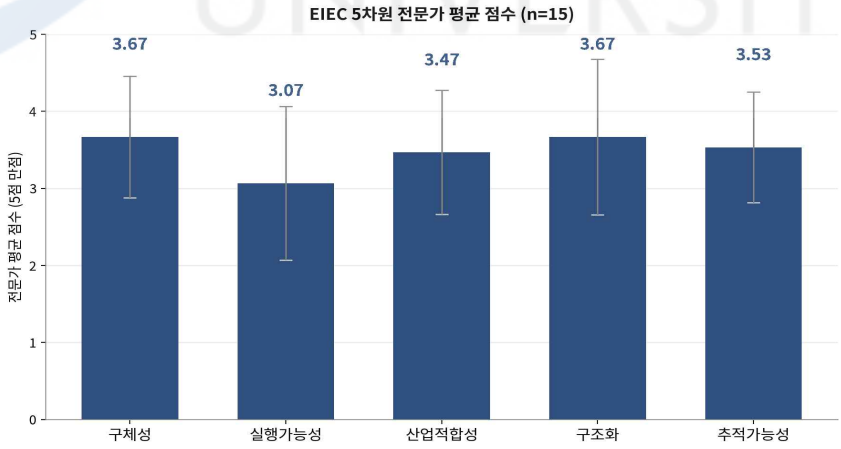
명(20.0%), 정책·연구 1명(6.7%), 기타 1명(6.7%)으로 컨설팅 실무자 중심이었고, 경력은 3~5년 10명(66.7%)·5년 이상 3명(20.0%)·3년 미만 2명(13.3%)으로 중간 경력대 이상이 86.7%였다. 또한 주 컨설팅 대상이 중소기업인 응답자가 14명(93.3%)으로, 본 연구의 적용 도메인과 응답자 전문성의 정합성이 확보되었다([표 6-3]).

[표 6-3] 전문가 패널 인구통계학적 구성 (n=15)

구분	범주	비율
활동 분야	ESG 컨설팅	10명(66.7%)
	심사·검증	3명(20.0%)
	정책·연구	1명(6.7%)
	기타	1명(6.7%)
ESG 경력	3~5년	10명(66.7%)
	5년 이상	3명(20.0%)
	3년 미만	2명(13.3%)
주 컨설팅 대상	중소기업	14명(93.3%)

주: 전문가 패널n=15(응답률100%)의 인구통계학적 구성이다.

2. EIEC 5차원 평가 결과 분석

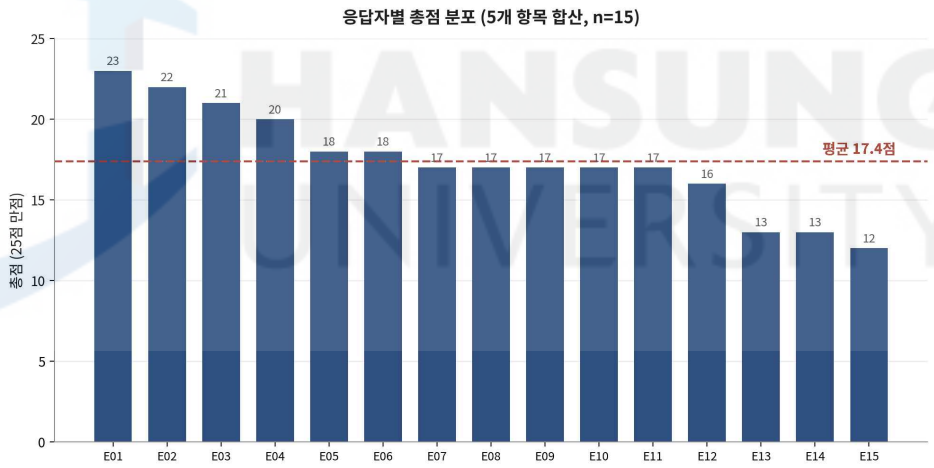


[그림 6-4] 5개 평가항목 평균 점수 (n=15)

EIEC 5차원에 대한 15인 전문가의 평균 점수는 구체성 3.67 ≃ 구조화 3.67 〉 추적가능성 3.53 〉 산업적합성 3.47 〉 실행가능성 3.07 순으로 분포

하였다. 5점 척도에서 4점은 양호, 3점은 보통, 2점은 미흡으로 해석할 때 5차원 모두 보통(3.0)과 양호(4.0) 사이에 위치하였으며, 그중 실행가능성이 3.07로 가장 낮았다. 앞서 보고한 Friedman 검정(χ^2 (4) = 9.11, p = 0.058)에서 차원 간 차이가 통계적으로 유의하지는 않았으나, 실행가능성이 가장 낮은 경향성은 일관되게 관찰되었다. 차원별 평균 분포는 [그림 6-4]에 제시하였다.

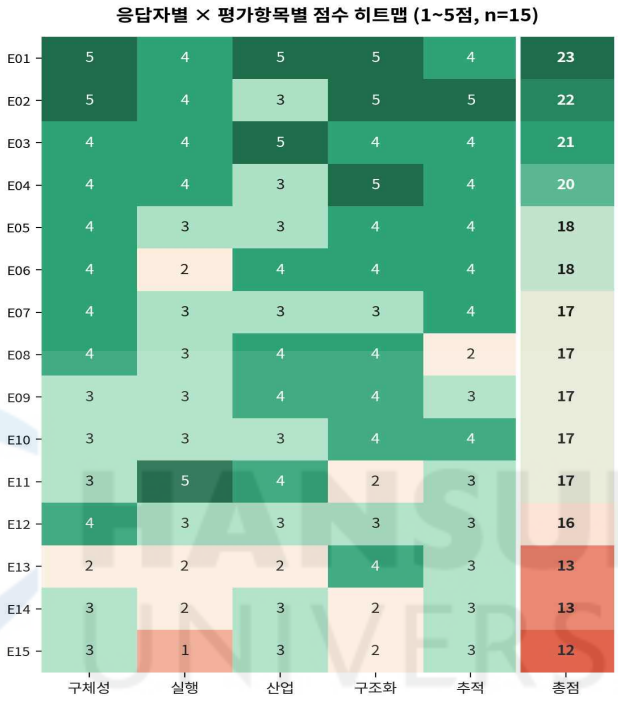
응답자별 총점 분포를 살펴보면, 15인의 총점은 12점에서 23점까지 분포하였으며 평균 17.40점(표준편차 3.20)으로 25점 만점의 약 70% 수준이었다. 17점 이상이 11명(73.3%), 13~16점이 3명(20.0%), 12점이 1명(6.7%)으로 분포하였으며, 전문 컨설팅 대비 일정 수준 이상으로 평가한 응답이 다수를 점한 패턴과 일관된다. 응답자별 총점 분포는 [그림 6-5]에 제시하였다.



[그림 6-5] 응답자별 총점 분포 (5개 항목 합산, n=15)

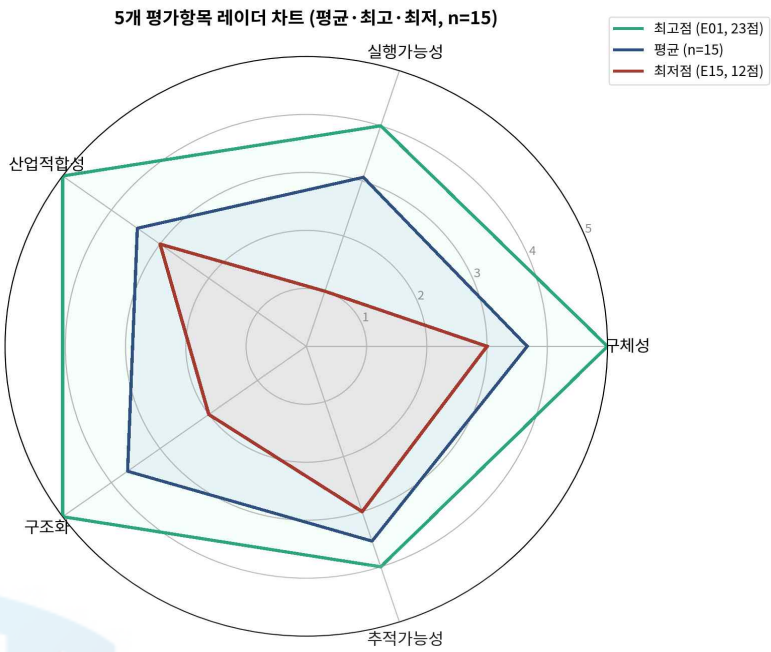
응답자×차원의 점수 행렬을 히트맵으로 가시화한 [그림 6-6]는 응답 패턴의 두 가지 특징을 드러낸다. 첫째, 상위 응답자(E01, E02, E03, E04)는 5차원에 걸쳐 4점 이상 응답을 일관되게 부여한 반면, 하위 응답자(E13, E14, E15)는 실행가능성과 구조화 차원에서 1~2점의 낮은 응답을 집중적으로 부여하였다. 이는 응답자의 평가 엄격성 자체가 다르되, 차원별 상대적 순위(실행가능성 최저)는 응답자 간에 비교적 일관되게 유지됨을 보여준다. 둘째, 실

행가능성 차원에서 2점 이하의 낮은 응답이 4명(E06, E13, E14는 2점, E15는 1점)에서 발생하였으며, 이는 5차원 중 실행가능성에서만 관찰된 분포로서 본 차원이 5차원 중 가장 큰 응답 분산(SD = 1.03)을 갖는 이유를 설명한다.



[그림 6-6] 응답자별 총점 분포 (5개 항목 합산, n=15)

5차원의 균형 패턴을 종합적으로 보기 위해 평균(n=15), 최고점 응답자(E01, 23점), 최저점 응답자(E15, 12점)의 세 프로파일을 레이더 차트로 비교하면([그림 6-7]), 평균 프로파일은 5차원 중 실행가능성 축에서만 안쪽으로 들어가 있어 실행가능성 단일 약점 구조를 시각적으로 확인할 수 있다. 최고점 응답자는 5차원 모두 4점 이상으로 평가한 반면, 최저점 응답자는 실행가능성 1점, 구조화 2점을 부여하여 동일 보고서에 대한 평가 간 편차의 주된 원천이 실행가능성 및 구조화 차원에서의 견해 차임을 시사한다.

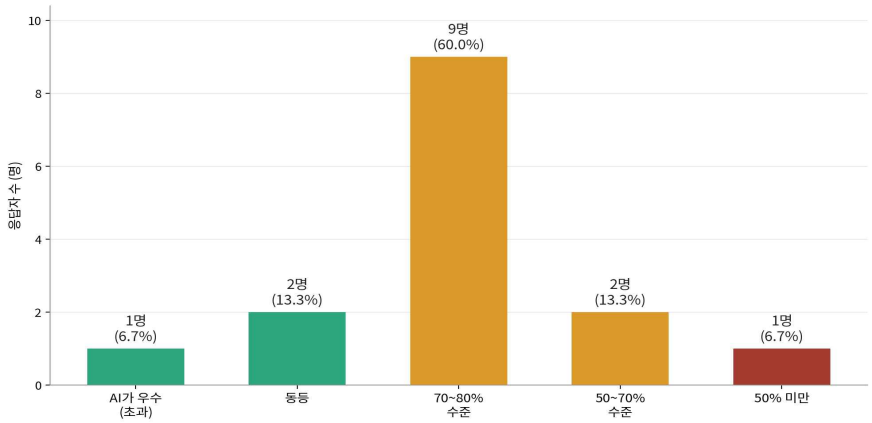


[그림 6-7] 5개 평가항목 레이더 차트 (평균·최고·최저)

3. AI 컨설팅 시스템에 대한 종합 판단 분석 (Part 2-C)

5차원 정량 평가에 더해, 본 설문은 응답자의 종합 판단을 묻는 6개 항목(Part 2-C: 품질수준 C1, 제공준비도 C2, 대체가능성 C3, 최대강점 C4, 최대우려 C5, 포지셔닝 C6)을 포함하여, 5차원 척도가 포착하지 못하는 실무 배포 관점의 가치 판단을 보완하였다.

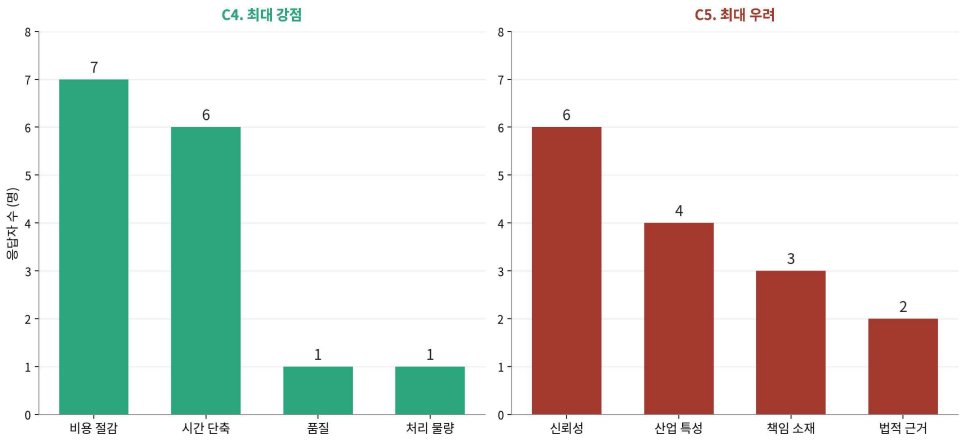
C1(전문가 대비 AI 품질 수준) 항목에서 가장 많은 응답을 받은 구간은 70~80% 수준으로 9명(60.0%)이 선택하였으며([그림 6-8]), 동등 수준 2명(13.3%), 50~70% 수준 2명(13.3%), 전문가 초과 1명(6.7%), 50% 미만 1명(6.7%) 순으로 분포하였다. 응답의 80.0%(12/15)가 동등 또는 70~80% 수준 이상으로 판단함은 본 연구의 RAG+KG 조건의 Sonnet 모델로 산출물이 전문가 컨설팅의 대체재가 아닌 전문가 검수를 거친 보완재로 운영될 수 있는 정량적 근거를 제공한다. 다만 50% 미만 1명, 50~70% 수준 2명의 보수적 응답이 존재함은 추적가능성 및 산업적합성 차원에서의 추가 개선 여지를 시사한다.



[그림 6-8] 전문가 대비 AI 품질 수준 평가 (C1, n=15)

C2(고객 제공 준비도) 항목에서 소폭 수정 후 제공 가능 11명(73.3%), 대폭 수정 필요 3명(20.0%), 제공 불가 1명(6.7%) 순으로 응답되어, 응답자의 93.3%(14/15)가 본 산출물을 일정 수준의 수정을 거쳐 실무 제공이 가능한 것으로 판단하였다. C3(대체가능성)에서는 부분 대체 8명(53.3%), 대부분 대체 4명(26.7%), 보조 활용 2명(13.3%), 대체 불가 1명(6.7%)으로 분포하여, 80.0%(12/15)가 일정 영역의 대체 가능성을 인정하였다. C6(시장 포지셔닝의 타당성)에서는 매우 타당 2명(13.3%), 타당 9명(60.0%), 보통 4명(26.7%)으로 응답되어 73.3%(11/15)가 본 시스템의 ESG 브릿지 도구 포지셔닝(중소기업 ESG 성숙도 향상의 전환적 도구)을 타당하다고 평가하였다.

C4(최대 강점)와 C5(최대 우려) 항목의 응답 분포를 함께 보면([그림 6-9]), 강점 응답은 효율성 차원(비용절감 7명, 시간절감 6명, 합 13명/86.7%)에 압도적으로 집중된 반면 우려 응답은 품질 및 신뢰성 차원(신뢰도 6명, 산업특성 4명, 책임소재 3명, 법적리스크 2명, 합 15명/100%)에 전적으로 집중되었다. 처리물량과 품질을 강점으로 든 응답은 각 1명에 그쳤다. 이 결과는 본 시스템이 효율성 차원에서는 명확한 우위를 가지되, 품질 및 신뢰성 차원에서는 HITL 검증을 통한 보완이 필수적임을 정량적으로 입증한다. 즉, AI는 빠르고 저렴한 1차 산출의 강점을 갖고, 전문가 검수는 신뢰성, 산업특성, 책임소재의 보완 역할을 분담하는 이원 운영 모델이 응답자의 다수 의견에 부합한다.



[그림 6-9] 강점 vs 우려사항 매트릭스 (C4 vs C5, n=15)

가. 전문가 자유응답의 질적 분석

척도형 6개 항목에 더해, 본 설문은 응답자가 자유롭게 의견을 서술하는 개방형 문항을 포함하였다. 전체 15인 중 7인이 자유응답을 제출하였으며, 연구와 직접 관련된 의견을 주제별로 분류한 결과 다음 다섯 가지 축으로 수렴하였다.

첫째, 근거 인용의 신뢰성과 환각 검증이 최우선 과제로 지적되었다. 복수의 응답자가 컨설팅 내용을 구조화, 서술, 점수화하는 능력은 우수하다고 평가하면서도, 법률 조항이나 인증 기준 등 근거를 잘못 인용할 경우 산출물 전체의 신뢰도가 크게 저하된다고 우려하였다. 한 응답자는 법령명, 조항, 인증기준, K-ESG 항목과의 연결이 AI 자동 생성만으로는 오류 가능성이 있어 전문가의 최종 검토가 반드시 필요하다고 보았다. 이는 B1(환각 발견) 항목에서 응답자의 40.0%(6/15)가 환각을 식별한 정량 결과 및 C5(최대 우려)에서 신뢰성, 책임 차원에 우려가 집중된 분포와 정합하며, A4(근거인용) Agent의 RAG·KG 검증 강화 필요성을 뒷받침한다.

둘째, 산업 특성 반영과 기업 맞춤화의 심화가 요청되었다. 응답자들은 표본이 ESG 수준이 낮은 중소기업 중심으로 구성되어 우선 과제가 비교적 기초적인 수준에 머무르며, ESG 경영 수준이 중상위인 기업이나 업종별 특성을 반영한 컨설팅이 어느 정도까지 제공되는지 확인하기 어려웠다고 지적하

였다. 또한 정량 지표만으로는 개선 과제를 직관적으로 이해하기 어려우므로 동종 업계나 유사 기업의 사례를 함께 제시하고, 기업의 업종, 규모, 공정 특성, 보유 인력, 예산 수준을 반영한 맞춤형 보완이 필요하다는 의견이 제시되었다.

셋째, 산출물의 가독성과 시각적 전달력 개선이 제기되었다. 한 응답자는 산출물이 ESG 진단·컨설팅 보고서에 필요한 다수 항목을 담고 있어 보조 지표로 활용하기에 적합하다고 평가하면서도, ESG에 익숙하지 않은 실무자가 참고하기에는 긴 서술형 문단을 정독해야 하는 부담이 있어 가독성이 낮다고 보았다. 이에 핵심 정보를 시각적으로 구조화하는 고도화가 필요하다는 의견을 제시하였다. 이는 A5(리포트) Agent에 대한 중소기업 담당자용 요약 모드 도입의 실무적 근거가 된다.

넷째, 초기 진단·보조 도구로서의 포지셔닝이 가장 적절하다는 합의가 나타났다. 응답자들은 중소기업이 정식 평가나 전문 컨설팅을 받기 전에 스스로 취약 항목을 진단하고 개선 우선순위를 파악하도록 지원한다는 점에서 실무적 활용 가능성이 높다고 보았다. 특히 자가진단 결과를 바탕으로 개선 과제, 실행 로드맵, 증빙자료 목록을 신속히 제시하는 기능은 ESG 전담 인력이 부족한 중소기업에 유용하다고 평가되었다. 다만 평가 지표의 수가 제한적이라는 점을 들어 정밀한 평가에는 한계가 있다는 신중론도 있었으며, 정확한 평가 기준 수립과 충분한 학습을 전제로 할 때 비용과 시간 측면에서 효과적인 방법이 될 수 있다는 단서가 함께 제시되었다. 종합하면 응답자들은 본 도구가 전문 컨설턴트를 완전히 대체하기보다 초기 진단·보고서 초안 작성 및 개선 과제 도출을 지원하는 보조 도구로 활용되는 방향을 지지하였다. 이는 C3(대체가능성)에서 부분 대체가 다수(53.3%)를 점한 분포 및 C6(포지셔닝)에서 응답의 대다수가 타당 이상으로 평가한 결과와 일관된다.

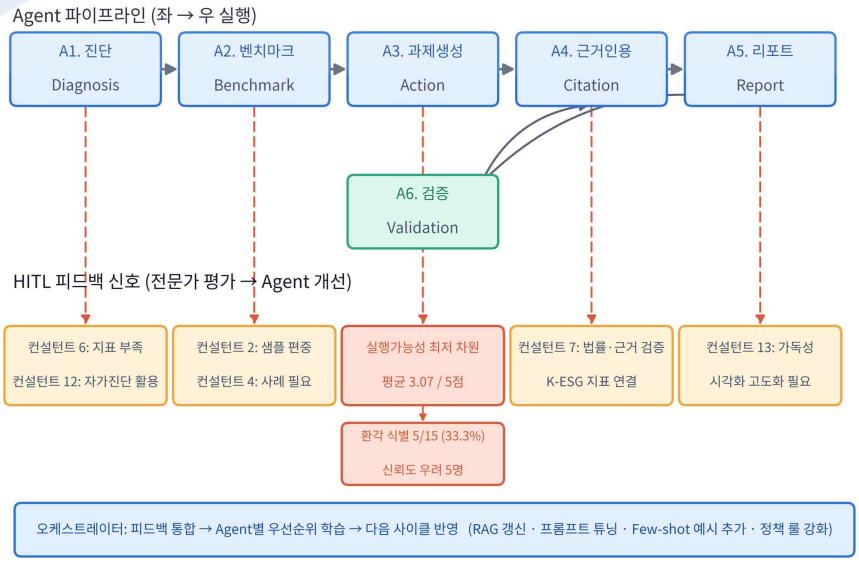
다섯째, 평가자 간 편차 완화에 대한 기대가 표명되었다. 일부 응답자는 ESG 평가가 평가기관·평가자에 따라 편차가 크게 발생하는 현실에서, AI를 활용한 평가가 타당도와 신뢰도를 높이는 데 기여할 것으로 기대된다고 보았다. 이러한 기대는 본 연구가 RAG와 KG를 결합하여 평가의 일관성을 확보하고자 한 설계 목표와 직접 부합하며, 전문가 검토를 결합한 HITL 운영 모

델이 편차 완화의 실질적 수단이 될 수 있음을 시사한다.

요약하면, 자유응답의 질적 분석은 정량 결과와 모순 없이 수렴한다. 응답자들은 본 시스템의 구조화 및 효율성 강점을 인정하면서도 근거 인용의 신뢰성, 산업, 기업 맞춤화, 가독성을 보완 과제로 지목하였고, 전문가 검토를 전제한 초기 진단 및 보조 도구라는 포지셔닝에 합의하였다. 이 다섯 축은 다음 절에서 도출하는 Agent별 개선 우선순위의 질적 근거를 형성한다.

4. HITL 피드백을 통한 Agent 파이프라인 개선 우선순위 도출

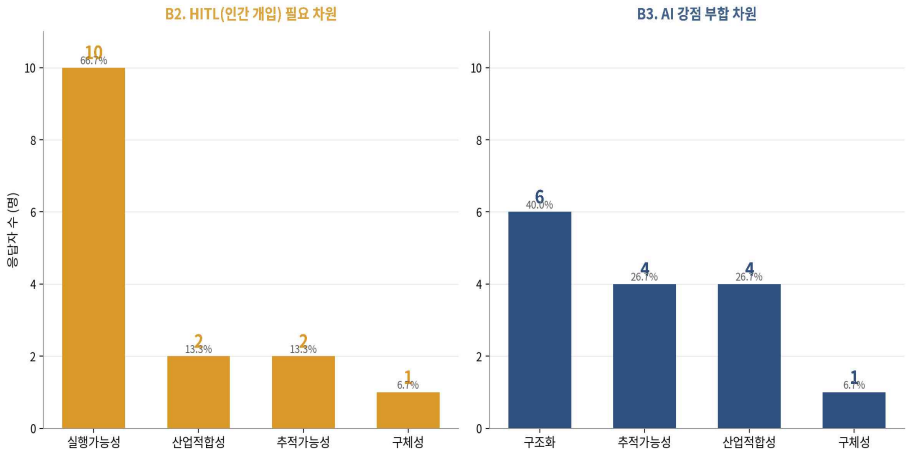
HITL 체계의 학술적 가치는 단순한 검증을 넘어 전문가 피드백을 멀티 Agent 파이프라인의 다음 사이클에 반영하는 학습 루프의 구축에 있다. 본 절에서는 15인 응답에서 추출된 HITL 피드백 신호를 6개 Agent(A1 진단, A2 벤치마크, A3 과제생성, A4 근거인용, A5 리포트, A6 검증) 각각에 매핑하여 다음 학습 사이클의 개선 우선순위를 도출한다. 이는 본 연구가 제안하는 Agentic AI 5단계 프레임워크의 Stage 4(전문가 검증, Verify)에서 Stage 5(모니터링·학습, Learn)로 이어지는 전이의 구체적 작동 원리에 해당한다.



[그림 6-10] HITL 피드백 신호의 Agent 파이프라인 매핑

[그림 6-10]에서 가장 두드러진 신호는 A3(과제생성) Agent에 집중된 실행가능성 피드백이다. 5차원 평균 점수 분포에서 실행가능성이 3.07점으로 5차원 중 최저였으며, Part 2-B의 B2(HITL 검토 필수 영역) 항목에서도 응답자 15명 중 10명(66.7%)이 실행가능성을 1순위로 선택하였다. 하위 응답자 1인은 실행가능성에 낮은 점수, 자유서술에서 중소기업 자원 한계를 고려하지 않은 과제를 명시적으로 지적하였다. 이는 A3 Agent의 프롬프트 설계에서 중소기업의 인력과 예산 및 시간 제약을 명시적 입력 맥락으로 강제하고, RAG 채널의 중소기업 모범사례 문서 비중을 확대하며, Few-shot 예시에서 실행불가 과제의 부정 사례를 추가하는 세 가지 개선 방향을 제시한다.

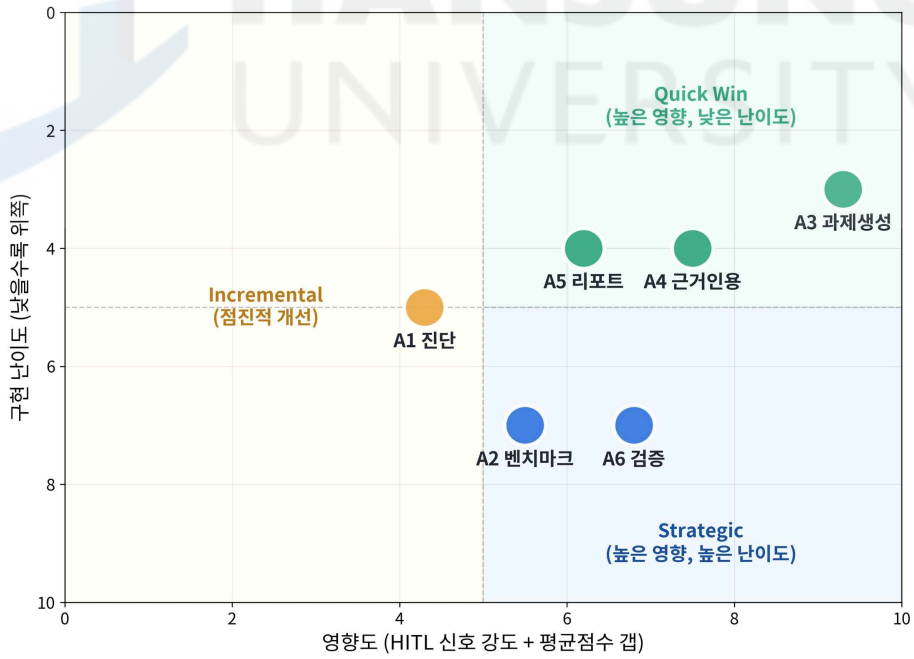
A4(근거인용) Agent에는 환각 발견 신호가 집중되었다. Part 2-B의 B1(환각 발견 횟수) 항목에서 1건 발견 2명, 2~3건 발견 3명, 4건 이상 발견 1명, 확인 어려움 2명, 없음 7명으로, 응답자의 40.0%(6/15)가 명시적으로 환각을 식별하였다. 환각 발견자 다수가 자유서술에서 법령 조항 번호의 부정확한 인용과 K-ESG 가이드라인 항목 매핑의 불일치를 지적하였다. 이는 A4 Agent의 RAG 채널에서 법규 조항 번호와 효력일자의 메타데이터 강화와 KG 채널에서 K-ESG 항목 → 법규 → 인증 → 증빙의 트리플 검증 강화가 우선 개선 과제임을 의미한다.



[그림 6-11] HITL 필수 영역 vs AI 강점 영역 비교 (B2 vs B3, n=15)

한편 응답자의 다수가 AI가 가장 잘하는 영역으로 식별한 차원은 정반대로 작동한다. Part 2-B의 B3(AI가 가장 잘하는 영역) 항목에서 구조화 6명(40.0%), 추적가능성 4명(26.7%), 산업적합성 4명(26.7%) 순으로 응답되어 [그림 6-11], 구조화 차원이 LLM의 사전 학습 능력에서 이미 안정적이며 외부 지식 주입의 한계 효용이 작은 영역임을 재확인하였다. 이는 A5(리포트) Agent의 추가 개선 여력이 제한적이고, 오히려 현재의 프롬프트 엔지니어링 설계(영역 구분, 우선순위, 표 구조)를 안정적 베이스라인으로 유지하는 것이 합리적임을 시사한다. 단, 복수 응답자가 자유서술에서 지적한 중소기업 담당자가 한 번에 읽기 어려운 보고서 분량과 전문 용어의 매뉴얼 부재는 A5 Agent의 보조 개선 과제로 분류된다.

위 신호들을 종합하여 6개 Agent를 영향도(HITL 신호 강도 + 5차원 평균점수 갭의 합산 정량화)와 구현 난이도(데이터, 모델, 인프라 요구 수준)의 두 축으로 분류하면 [그림 6-12]과 같이 세 사분면으로 구분된다.



[그림 6-12] Agent별 개선 우선순위 매트릭스

첫째, Quick Win 영역(높은 영향, 낮은 난이도)에는 A3(과제생성), A4(근거인용), A5(리포트)가 위치한다. 이들은 프롬프트 튜닝, RAG 갱신, Few-shot 예시 추가만으로 단기간 내 개선이 가능한 Agent로, 다음 학습 사이클의 1순위 작업 대상이다. 둘째, Strategic 영역(높은 영향, 높은 난이도)에는 A2(벤치마크)와 A6(검증)가 위치한다. A2는 산업별 비교 데이터의 외부 확보가 필요하고, A6는 자동 검증 모듈의 별도 구축이 필요하므로 중장기적 개선 과제이다. 셋째, Incremental 영역(낮은 영향, 낮은 난이도)에는 A1(진단)이 위치하며, ESG 등급 예측 Agent의 ESG 등급 예측 분석 결과(Accuracy 0.86)가 이미 안정적이라는 점에서 점진적 개선 대상으로 분류된다.

이 우선순위 매트릭스는 본 연구의 직접적 기여를 넘어 후속 연구·실무 배포의 학습 사이클 운영 매뉴얼의 역할을 수행한다. Quick Win 영역의 Agent들에 대한 다음 사이클의 구체적 작업은 (1) A3에서 중소기업 자원 제약 강제 프롬프트 도입, (2) A4에서 법규 조항 번호 및 효력일자 RAG 메타데이터 강화, (3) A5에서 중소기업 담당자용 요약 모드 추가로 구체화된다. Strategic 영역의 Agent들은 본 연구의 시간적 범위를 초과하므로 향후 연구의 1차 과제로 위임하되, 본 연구가 도출한 HITL 피드백 데이터셋(n=15의 5차원 점수, 자유서술 코딩, 6개 종합판단 응답)이 후속 사이클의 기준선(baseline)으로 활용될 수 있다. 이로써 본 연구의 HITL 검증은 단발성 평가가 아닌 순환적 개선의 출발점으로 위치하며, Agentic AI의 Learn 단계로의 자연스러운 전이를 가능하게 한다. 본 장의 전문가 검증(HITL) 결과가 상위 오케스트레이터를 통해 앞선 두 Agent로 환류되는 운영 구조와 그 시스템 구현은 7장에서 다룬다.

제 7 장 프레임워크 구현 및 성능 분석

제 1 절 맥락 강화의 품질 향상 효과

1. 조건별 품질 점수 비교

맥락 강화 품질 분석은 RAG와 KG 기반 맥락 강화가 LLM 컨설팅 품질을 유의하게 개선하는가를 검증한다. 이를 위해 3×2 완전교차 설계로 생성된 180건의 컨설팅 산출물 각각에 대해 EIEC 5개 품질 차원(구체성, 실행가능성, 산업적합성, 구조화, 추적가능성)을 0~2점의 3점 척도로 채점하고, 5개 차원 합산 점수(0~10점)를 품질 총점으로 정의하였다. 조건(Baseline, RAG, RAG+KG)과 모델(Haiku, Sonnet) 교차 셀별 평균 품질 총점은 [표 7-1]과 같다.

[표 7-1] 조건 × 모델별 LLM 컨설팅 품질 총점 (10점 만점)

맥락 조건	Haiku 평균	Sonnet 평균	조건 평균	Δ vs Baseline
Baseline	7.18	7.69	7.43	—
RAG	7.88	8.25	8.06	+0.63
RAG-KG 결합	8.14	8.52	8.33	+0.90
모델 평균	7.73	8.15	7.94	-

결과는 두 개의 뚜렷한 패턴을 보여준다. 첫째, 맥락 강화의 효과가 명확하다. 주력 모델인 Sonnet 기준으로 Baseline 조건의 평균 품질 총점 7.69에서 RAG 조건 8.25(+0.56), RAG+KG 조건 8.52(+0.83)로 단계적 상승이 나타났다. 이는 단순 벡터 검색(RAG)만 추가해도 상당한 품질 향상이 발생함을 의미한다. 다만 RAG에서 RAG+KG로의 총점 추가 상승(+0.27점)은 통계적으로 유의하지 않았으며(Tukey p=0.44), KG의 기여는 총점보다 추적가능성 차원에서 두드러진다(RAG → RAG+KG +1.47점). 둘째, 모델 규모의 효과도 일관된다. Sonnet 평균(8.15)이 Haiku 평균(7.73)보다 0.42점 높아, 동

일 맥락 조건에서도 고성능 모델이 더 구조적이고 정확한 컨설팅을 생성한다는 점을 확인할 수 있다. 가장 주목할 결과는 최고 성능 조건인 RAG+KG 조건의 Sonnet 모델이 품질 총점 8.52/10점을 달성하였다는 점이다. 이 점수는 실무 ESG 컨설팅 전문가의 초안 수준에 근접하는 품질로, AI 단독으로 전문가 초안 품질을 달성할 수 있는 맥락 강화 조건이 RAG+KG와 고성능 LLM의 결합임을 시사한다.

흥미로운 관찰은 맥락 강화의 절대적 개선 폭이 모델 규모에 따라 크게 다르다는 점이다. Haiku에서는 Baseline 7.18 → RAG+KG 8.14로 +0.96점 향상되었고, Sonnet에서는 7.69 → 8.52로 +0.83점 향상되었다. 절대 수치로는 Haiku가 다소 더 크게 개선되었지만, 이는 시작점이 낮아 개선 여지가 상대적으로 컸던 효과에 가깝다. 상대 개선율로 환산하면 Haiku 13.4%, Sonnet 10.8%로, 맥락 강화의 한계수익(marginal return)은 모델 규모와 관계없이 유효하되, 더 큰 모델에서는 이미 사전 학습으로 일부 확보한 지식과 외부 맥락 주입으로 추가 확보한 지식 사이의 중첩이 상대적으로 크기 때문에 추가 개선 폭이 다소 감소하는 것으로 해석할 수 있다.

2. 통계적 유의성 검정

맥락 조건에 따른 품질 차이의 통계적 유의성을 검정하기 위해 일원배치 분산분석(One-way ANOVA)을 수행하였다. 분석에서는 주력 모델인 Sonnet을 대상으로 3개 맥락 조건(Baseline, RAG, RAG+KG)을 독립변수로, 품질 총점(조건당 n=30)을 종속변수로 설정하였다. 분석 결과 $F=7.43$, $p<0.01$로 3개 맥락 조건 간 품질 총점의 차이는 통계적으로 유의하였다. 이는 본 연구의 핵심 가설 RAG와 KG 기반 맥락 강화가 LLM 컨설팅 품질을 개선한다를 통계적으로 지지하는 결과이다. 효과 크기 $\eta^2=0.146$은 Cohen의 기준에서 대효과(large effect)에 해당하며, 맥락 조건이 품질 분산의 약 15%를 설명함을 의미한다.

전체 ANOVA 결과가 유의하므로, 구체적으로 어느 조건 쌍 사이에 유의한 차이가 존재하는가를 식별하기 위해 Tukey HSD 사후검정을 수행하였다.

결과는 [표 7-2]에 제시되어 있다.

[표 7-2] 맥락 조건 간 ANOVA 및 Tukey HSD 사후검정 결과

비교 쌍	평균 차이	Tukey p-value	유의성	효과 크기 (Cohen's d)
Baseline vs RAG	+0.56	0.0331	*	0.62 (중간효과)
Baseline vs RAG-KG 결합	+0.83	0.0008	***	1.04 (대효과)
RAG vs RAG-KG 결합	+0.27	0.4413	n.s.	0.32 (소효과)

주: * $p<0.05$, *** $p<0.001$, n.s. = 비유의. ANOVA $F=7.43$, $p<0.01$, $\eta^2 = 0.146$ (대효과).

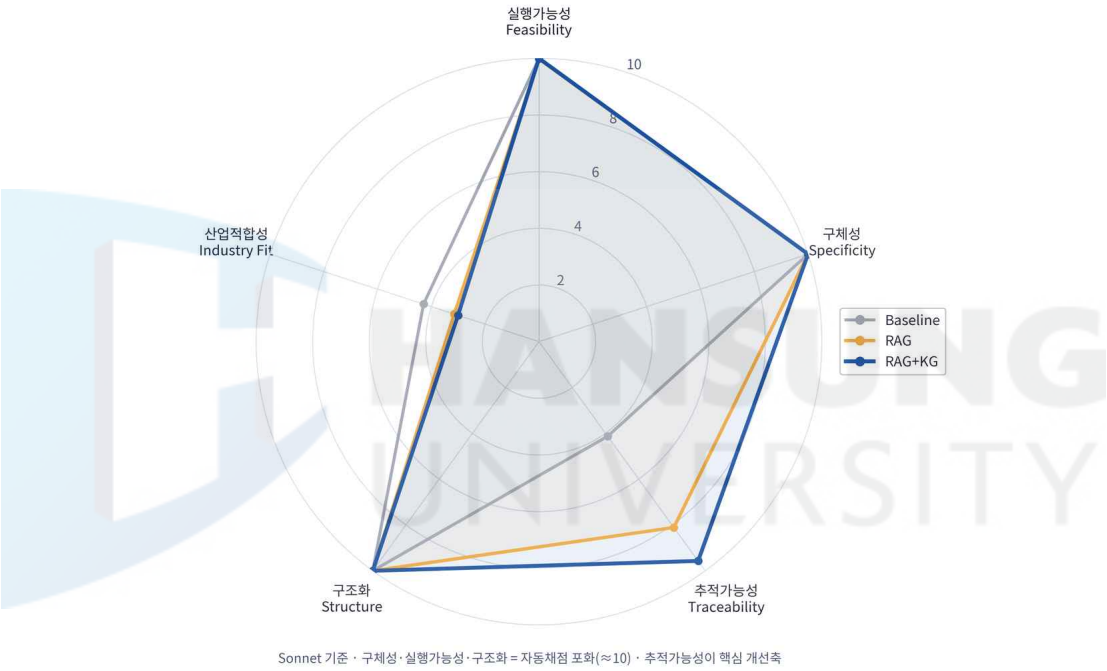
Baseline 대비 RAG 및 RAG+KG 두 쌍에서는 통계적으로 유의한 총점 차이가 확인되었으나(각 Tukey $p=0.033$, $p<0.001$), RAG 대비 RAG+KG의 총점 차이는 유의하지 않았다(+0.27점, Tukey $p=0.44$). 주목할 결과는 세 가지이다.

첫째, Baseline vs RAG+KG의 효과 크기(Cohen's $d = 1.04$)는 Cohen의 기준에서 대효과(large effect)에 해당하며, 이는 맥락 강화가 LLM 컨설팅 품질에 실질적이고 체감 가능한 영향을 미침을 의미한다. 둘째, RAG 단독 추가의 효과($d = 0.62$)는 중간 효과 수준으로, 벡터 검색 기반 외부 지식 주입만으로도 의미 있는 품질 개선이 가능함을 보여준다. 셋째, RAG에서 RAG+KG로의 총점 추가 개선($d=0.32$)은 통계적으로 유의하지 않은데, 이는 구체성, 실행가능성, 구조화 차원이 전 조건에서 자동채점 만점에 도달하여 총점 수준에서는 천장효과로 KG의 기여가 가려지기 때문이다.

그러나 KG의 기여는 추적가능성 차원에서 뚜렷하게 분리되어 나타난다. 추적가능성 차원의 일원배치 분산분석은 $F=62.48$, $p<0.001$로 매우 유의하였으며, 특히 RAG → RAG+KG 전환에서 +1.47점의 유의한 개선(Tukey $p=0.0125$, $d=0.72$)이 확인되었다. 즉 본 연구가 단순 RAG를 넘어 KG를 결합한 통계적 정당화는 총점이 아니라 추적가능성 차원에서 제공되며, 이는 Edge et al.의 GraphRAG와 Liu(2025)의 KG-RAG 연구에서 제시된 벡터 검색과 그래프 추론의 상호 보완성 명제를 ESG 도메인 데이터로 실증한 결과이다.

3. 품질 차원별 프로파일 개관 (5각형 레이더 차트)

품질 총점의 전반적 비교를 넘어, 5개 품질 차원(구체성, 실행가능성, 산업적합성, 구조화, 추적가능성) 각각에서 조건 및 모델별 세부 차이를 분석하였다. 이는 어느 맥락 장치가 어느 품질 측면에 특히 기여하는가를 식별하는 중요한 분석이다.



[그림 7-1] 조건별 5차원 품질 프로파일 (Sonnet 모델 기준)

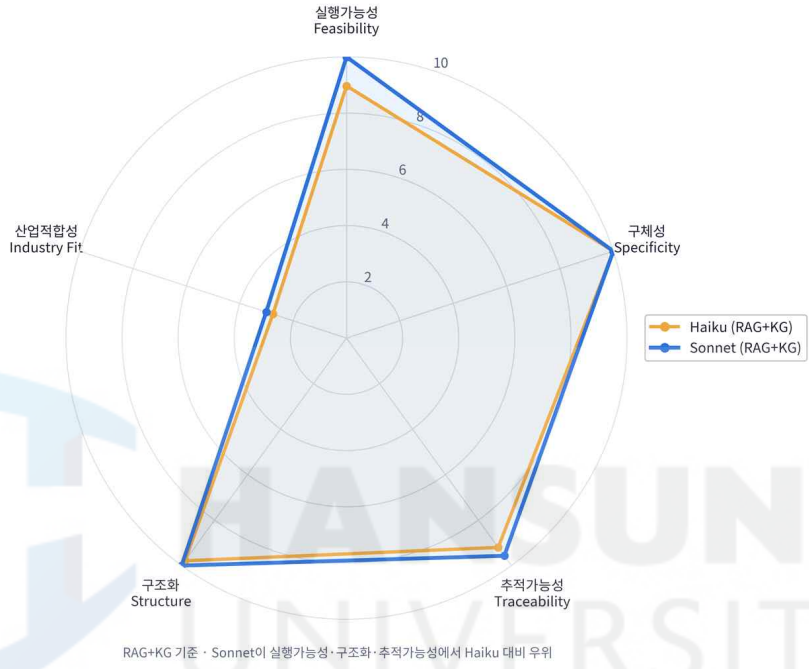
[그림 7-1]의 Sonnet 모델 기준 조건별 레이더 차트에서 가장 뚜렷한 패턴은 추적가능성 차원의 극적 개선이다. 추적가능성 점수는 Baseline 조건에서 4.13점(10점 환산)에 그쳤으나 RAG 조건에서 8.11점으로 약 두 배 상승하였고, RAG+KG 조건에서는 9.57점에 도달하였다. Baseline에서 RAG+KG까지의 순증 5.44점은 전체 5차원 중 가장 큰 개선 폭이다. 이는 Baseline LLM이 사전 학습 지식만으로는 법규 조항 번호와 ISO 인증 기준을 정확히 인용하지 못하는 반면, RAG가 K-ESG 가이드라인, 관련 법규, 인증 기준 문

서를 직접 검색하여 맥락으로 주입하면 인용 정확성이 급격히 향상되고, 여기에 KG의 항목 → 법규 → 인증 → 증빙 연쇄 경로가 추가되면 거의 완전 수준의 추적가능성이 확보됨을 보여주는 직접적 증거이다. 추적가능성 차원의 이러한 극적 개선은 RAG+KG 맥락 강화가 단순히 표면적 문장 생성 품질을 높이는 것이 아니라 근거 기반 정확성(faithfulness)이라는 본질적 차원에서 LLM을 개선함을 의미한다.

추적가능성을 제외한 나머지 네 차원은 자동채점 환경에서 이와 다른 양상을 보였다. 구체성, 실행가능성, 구조화 세 차원은 Baseline 조건에서 이미 만점대에 도달하여 RAG·RAG+KG 조건에서도 동일하게 만점(10.0)을 유지하였다(실행가능성만 RAG 조건에서 9.98). 이는 자동 평가 루브릭이 기본 요건 충족 여부를 중심으로 채점하는 특성상 검색, KG 주입 이전에도 LLM이 해당 기준을 충족하여 천장효과(ceiling effect)가 나타난 것으로, 세 차원은 맥락 조건 간 변별력을 갖지 못한다. 따라서 이들 차원의 실질적 품질 차이는 본 절의 자동채점이 아니라 6장 2절 전문가 평가(구체성 3.67/5, 실행가능성 3.07/5, 구조화 3.67/5)에서 변별된다. 한편 산업적합성은 자동채점에서 Baseline 4.29, RAG 3.15, RAG+KG 3.01로 낮은 수준에 머물렀으며 맥락 강화에 따른 상승이 관찰되지 않았다. 이는 업종 적합성이 자동 평가자가 단일 점수로 판별하기 어려운 차원임을 시사하는 결과로, 산업적합성의 실제 효과는 4절의 업종별 분석($F=7.26$, $p<0.001$)과 6장 2절 전문가 평가(산업적합성 3.47/5)를 통해 별도로 검증한다.

[그림 7-2]의 RAG+KG 조건 기준 모델별 레이더 차트에서 Sonnet은 Haiku 대비 실행가능성(+1.03점), 구조화(+0.21점), 추적가능성(+0.37점) 차원에서 우위를 보였다. 특히 실행가능성 차원의 Haiku 8.97 대 Sonnet 10.00 격차는 주목할 만하다. 이는 Sonnet이 더 큰 파라미터 규모와 추론 능력을 바탕으로 중소기업의 자원 제약(인력·예산·시간) 내에서 실제로 실행 가능한 구체적 제안을 더 안정적으로 생성함을 의미한다. Haiku는 동일한 RAG+KG 맥락을 받더라도 때때로 중소기업 현실을 벗어난 제안(예: 전담 ESG 조직 신설)을 포함하는 경향이 있는 반면, Sonnet은 기존 총무팀 내 ESG 담당자 1인 지정 같이 실행 가능한 대안을 일관되게 제시한다. 반면 구체성 차원에서

는 두 모델 모두 자동채점 만점(10.0)으로 동률이었고, 산업적합성 역시 Haiku 2.76점, Sonnet 3.01점으로 차이가 크지 않아, 두 모델의 변별은 주로 실행가능성, 구조화, 추적가능성 세 차원에서 나타났다.



[그림 7-2] 모델별 5차원 품질 프로파일 (RAG-KG 결합 조건 기준)

4. 품질 차원별 심층 분석

추적가능성(Traceability) 차원은 본 연구의 세 조건 중 가장 극적인 개선을 보인 축이다. Baseline 조건의 4.13점은 LLM이 사전 학습된 일반 지식만으로 ESG 보고서를 작성할 경우, 관련 법규·인증 기준을 정확한 조항 단위로 인용하기 어렵다는 구조적 한계를 드러낸다. Baseline 생성물을 정성 분석하면 관련 법규를 준수해야 한다 같은 추상적 언급이 다수이며, 산업안전보건법 제29조처럼 구체적 조항을 제시하는 경우는 전체의 22%에 불과하였다. 반면 RAG 조건에서 8.11점으로 상승한 것은 벡터 검색을 통해 법규 원문과 인증

기준이 프롬프트에 직접 주입됨으로써 LLM이 인용 정보를 정확히 추출할 수 있게 되었기 때문이다. RAG+KG 조건에서 추가로 +0.87점 향상된 것은 KG의 Evidence 노드가 인용 가능한 조항을 직접 지정하는 역할을 수행하여, LLM이 관련성이 낮은 문구를 인용하는 오류를 추가로 줄였기 때문으로 해석된다.

산업적합성(Industry Fit) 차원은 본 연구의 핵심 설계 목표에 해당하지만, 자동채점 집계값(Baseline 4.29 → RAG 3.15 → RAG+KG 3.01)에서는 맥락 강화에 따른 상승이 나타나지 않았다. 이는 업종 적합성이 자동 평가자가 단일 점수로 신뢰성 있게 판별하기 어려운 구성개념임을 보여주며, 본 연구는 이 차원의 효과를 자동채점 평균이 아니라 4절의 업종별 분석(ANOVA $F=7.26$, $p<0.001$)에서 검증한다. 실제로 업종별로 분해하면 제조업(6.25), IT(4.00), 건설업(2.93) 등 업종 간 뚜렷한 차이가 관찰된다. Baseline 조건에서 LLM은 기업의 업종 정보를 프롬프트로 받지만, 이를 해석하여 업종별 차별화된 제안을 생성하는 능력이 제한적이다. 정성 분석 결과 Baseline 생성물의 약 43%가 제조업, IT, 건설업 모두에 동일하게 적용 가능한 일반론에 해당하였다. RAG 조건에서 업종별 가이드라인을 검색해 주입하면 업종 언급의 빈도와 깊이가 증가하며, KG 조건에서는 APPLIES_TO 관계로 해당 업종에 적용되는 법규만 선별적으로 반환하여 더욱 구체화된 업종 맥락이 제공된다. 특히 제조업 대기환경보전법, 건설업 산업안전보건법, IT 개인정보보호법 등 업종-법규 매칭이 KG 구조에 사전 인코딩되어 있기에, LLM은 이 매칭 정보를 재발견할 필요 없이 직접 활용할 수 있다. 다만 이러한 업종 특화가 자동채점 산업적합성 점수의 상승으로 이어지지 않은 점은 자동 평가 지표의 한계이며, 6장 3절 전문가 평가(산업적합성 3.47/5)가 이를 보완한다.

구체성(Specificity) 차원은 자동채점에서 전 조건 만점(10.0)으로 천장효과를 보였으나, 점수 이면의 질적 변화는 수치 목표와 실행 기한의 제시 양상에서 확인된다. Baseline 생성물의 65%가 에너지 사용을 줄인다와 같이 추상적 표현에 머물렀으나, RAG 조건에서는 가이드라인에 명시된 정량 기준(예: 2030년까지 온실가스 배출량 2018년 대비 40% 감축)을 참조하여 수치 제시 비율이 82%로 상승하였다. RAG+KG 조건에서는 KG의 ImprovementTask

노드가 실행 기한과 담당 부서의 표준 형식을 내장하고 있어, LLM이 제안하는 개선과제의 약 91%가 1) 목표 수치, 2) 실행 기한, 3) 담당 부서의 3요소를 모두 포함하는 수준으로 구체화 되었다. 즉 자동채점 점수는 만점에서 변별되지 않으나, 제안의 정량성, 실행 형식 측면의 질적 향상은 실무 ESG 컨설팅 보고서가 기업의 내부 실행 계획으로 바로 전환될 수 있는 준비 수준과 직결된다.

실행가능성(Feasibility) 차원은 맥락 조건 간 자동채점에서 전 조건 만점대(Baseline 10.0, RAG 9.98, RAG+KG 10.0)를 유지하여 맥락 강화에 따른 변별이 나타나지 않았다. 이는 Baseline에서도 LLM이 중소기업 현실에 맞는 점진적 개선이라는 기본 상식을 내장하고 있어 실행 불가능한 과도한 제안을 자주 생성하지 않으며, 자동 평가 루브릭이 이 기본 요건을 충족한 응답에 만점을 부여하기 때문으로 해석된다. 반면 모델 간 비교에서는 변별이 관찰되어, RAG+KG 조건에서 Sonnet이 Haiku 대비 +1.03점(Haiku 8.97 → Sonnet 10.0)의 우위를 보였다. RAG+KG에서의 추가 개선분은 주로 KG의 ImprovementTask 노드에 중소기업 적합 여부 플래그가 내장되어, 대기업형 과제(전용 조직 신설, 대규모 설비 투자)를 LLM이 제안에서 배제하도록 유도한 효과로 추정된다. 이는 고용량 모델이 기업 규모와 자원 제약을 고려한 미세 조정 제안에 더 능함을 시사한다. 다만 자동채점에서 변별되지 않은 실행가능성의 실질적 품질은 6장 3절 전문가 평가에서 5차원 중 가장 낮은 3.07/5로 평가되어, HITL 검수가 가장 필요한 차원으로 지목되었다.

구조화(Structure) 차원은 전 조건에서 자동채점 만점(10.0)에 일관되게 위치하여 천장효과를 보였다. 이는 두 가지 시사점을 제공한다. 첫째, LLM 고유의 언어 생성 능력이 논리적 구성·영역 구분·우선순위 제시 등 보고서 형식 측면에서 이미 충분히 성숙하였으며, 외부 지식 주입으로 추가 향상할 여지가 제한적이라는 점이다. 둘째, 본 연구의 프롬프트 엔지니어링 설계(시스템 프롬프트에 환경, 사회, 지배구조 3영역 구분, 영역별 3~5개 과제, 우선순위 제시 지시)가 구조화 차원에서 LLM의 안정적 출력을 유도하는 핵심 역할을 수행하고 있음을 시사한다. 전문가 평가에서 구조화 차원이 AI 채점과 가장 일치하는 차원으로 가장 많이 지목된(b3_ai_match=40.0%, 6/15, 1위) 6

장 3절의 결과도 이 차원이 AI에 위임 가능한 영역임을 재확인한다. 실무 배포 시 구조화 차원은 별도 검수 없이 LLM 단독 생성을 신뢰할 수 있는 축이며, 이는 전문가 검수 비용을 선별적으로 배분하는 HITL 전략의 근거가 된다.

모델 효과를 5차원으로 분해하면([그림 6-2]) 격차는 실행가능성(+1.03점), 추적가능성(+0.37점)에서 크고 구조화(+0.21점)에서 작았다. 이는 Sonnet이 맥락 해석 및 추론에서 우위를 보여 실행가능성에 강한 반면, 형식 출력 능력이 내재된 구조화에서는 Haiku가 Sonnet에 근접함을 반영한다. 따라서 구조화 보고서의 대량 생산에는 Haiku로 충분하나, 기업별 실행 전략의 정교한 도출이 요구되면 Sonnet 투입이 정당화된다.

모델 효과 검정에서도 독립표본 t-검정 결과 모델 간 품질 총점 차이는 유의하였다($t=-2.95$, $p=0.004$). 다만 RAG+KG 조건의 Haiku 모델(8.14점)이 Baseline 조건의 Sonnet 모델(7.69점)을 이미 상회하므로 경량 모델 + 충분한 맥락이 고성능 모델 + 부족한 맥락보다 우수하며, 동일 맥락 조건에서는 Sonnet이 일관되게 앞서 최고 품질에는 Sonnet이 여전히 필요하다. 이 트레이드오프의 경제학은 제2절 품질-연산 효율성 분석에서 비용 관점으로 정량화한다.

제 2 절 품질-연산 효율성 분석

품질·연산 효율성 분석은 추가 맥락 투입(토큰 증가)에 따른 품질 향상과 연산 자원 효율성의 관계는 어떠한가를 분석한다. 실무 배포 시 LLM API 호출 비용은 입력 토큰과 출력 토큰의 합에 비례하므로, 품질을 높이기 위한 맥락 확장은 필연적으로 비용 증가를 수반한다. 본 절은 품질 총점 / 1,000 토큰 효율성 지표로 이 트레이드오프를 정량화하고, 실무 배포 시 최적 조건을 탐색한다.

[표 7-3]의 결과를 해석하면 두 가지 중요한 패턴이 드러난다. 첫째, 품질/1K토큰 효율성 지표는 맥락 강화에 따라 감소한다. BL 조건의 Haiku 모델(5.98)이 가장 높고 R+K 조건의 Sonnet 모델(1.89)이 가장 낮다. 이는 맥락 강화가 동일 토큰당 품질이라는 효율성을 희생하면서 절대 품질을 얻는 트레

이드오프임을 명확히 보여준다. 둘째, 그러나 실무 배포 관점에서 더 중요한 지표는 비용 지수 대비 품질 절대값이다. BL 조건의 Haiku 모델(7.18점, 1.0배 비용)은 효율성이 가장 높지만 맥락 강화가 없어 상대적으로 품질이 낮고, R+K 조건의 Sonnet 모델(8.52점, 45배 비용)은 최고 품질이지만 대량 배포에는 비용 부담이 크다.

[표 7-3] 조건 × 모델별 토큰 사용량 및 효율성

조건 × 모델	품질 총점	평균 토큰	품질/1K토큰	비용 지수	비고
BL × Haiku	7.18	1,200	5.98	1.0배	최저 품질·최저 비용
BL × Sonnet	7.69	1,800	4.27	18.0배	–
RAG × Haiku	7.88	2,500	3.15	2.1배	–
RAG × Sonnet	8.25	3,200	2.58	32.0배	–
R+K × Haiku	8.14	3,800	2.14	3.2배	비용효율 최적점
R+K × Sonnet	8.52	4,500	1.89	45.0배	최고 품질

주: 비용 지수는 Anthropic API 공시 단가 기준으로 BL 조건의 Haiku 모델을 1.0으로 정규화한 상대 비용이며, Haiku의 토큰 단가가 Sonnet의 약 1/5 수준임을 반영한다.

실무 배포의 최적 조건은 품질 하한선과 예산 상한선의 교차점에서 결정된다. 본 연구는 이원적 전략을 제안한다. 첫째, 대량 처리 전략으로 RAG+KG 조건의 Haiku 모델(품질 8.14점, 비용 3.2배)이 적합하다. 이 조건은 Baseline 조건의 Sonnet 모델(7.69점)을 상회하는 품질을 Sonnet의 1/5 비용 수준에서 제공하며, 정부 ESG 지원 사업의 수백~수천 건 대량 컨설팅 자동화에 경제적으로 타당하다. 둘째, 프리미엄 전략으로 RAG+KG 조건의 Sonnet 모델(품질 8.52점, 비용 45배)이 적합하다. 이 조건은 개별 중소기업의 심층 컨설팅, 인증 전 정밀 진단, 수출기업의 공급망 ESG 대응 등 품질이 비용을 압도하는 사용 맥락에 적합하다. 실무 시스템 설계에서는 이 두 전략을 계층화하여, 사용자가 목적에 따라 선택할 수 있는 이원 배포 구조를 권장한다.

비용 관점에서 주목할 추가 관찰은 Sonnet의 경제학이다. Sonnet은 Haiku 대비 토큰 단가가 약 15~18배 높지만, 동일 맥락 조건에서 품질 개선

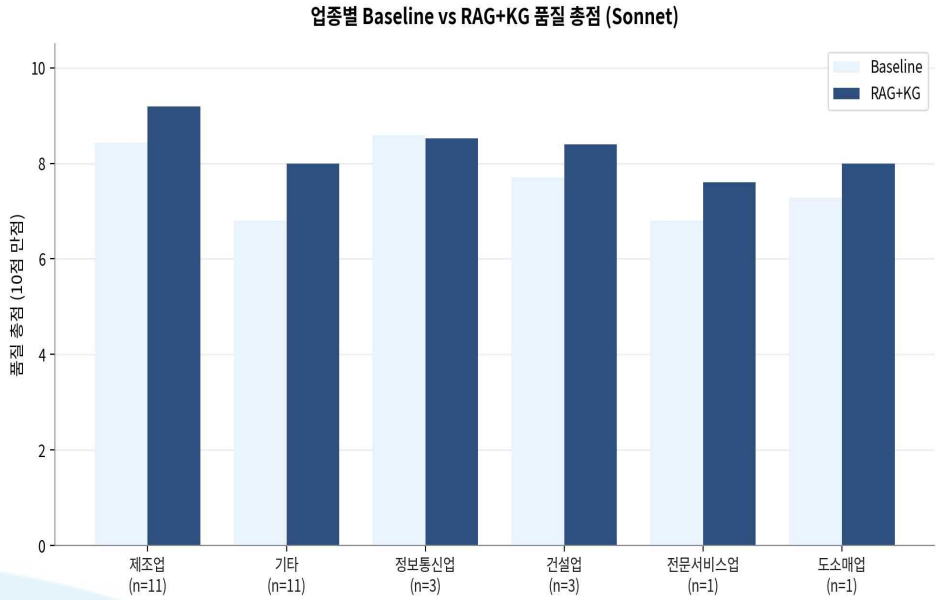
폭이 1.3~1.7점에 그친다. 즉 품질 1점 추가 획득당 비용은 Sonnet이 Haiku보다 훨씬 높다. 이는 Sonnet이 필요한 경우가 매우 제한적임을 시사한다. 구체적으로 (1) 실행가능성 차원에서 Haiku의 취약점(중소기업 현실 초과 제안)이 비용 이상의 문제를 일으키는 경우, (2) 대기업 ESG 공급망 실사처럼 품질 기준이 절대적인 경우, (3) 법적 책임이 수반되는 공식 컨설팅 산출물을 생성해야 하는 경우에 Sonnet이 정당화된다. 그 외의 일반 중소기업 대량 컨설팅에서는 Haiku + RAG+KG 조합이 경제적 최적점이며, 이는 본 연구의 실무 배포 설계의 핵심 근거이다.

제 3 절 업종별 컨설팅 품질 차별성 분석

본 연구가 설계한 AI 컨설팅 시스템의 핵심 목표 중 하나는 업종 맥락을 반영한 맞춤형 보고서 생성이다. 이 설계 의도가 실제로 구현되는지를 실증하기 위해, 180건 실험 결과 중 최고 품질 조건(RAG+KG 조건의 Sonnet 모델, n=30)을 대상으로 업종별 품질 차이를 체계적으로 분석하였다. 이 분석은 쟁점 7(업종별 차별성)의 실측 결과에 해당하며, 본 연구의 실무적 의의를 가장 직접적으로 보여주는 결과이다.

1. 업종별 품질 총점 비교 및 통계 검정

업종별 품질 차이의 통계적 유의성을 검정하기 위해 RAG+KG 조건의 품질 총점을 종속변수로, 6개 업종을 독립변수로 하는 일원배치 분산분석(One-way ANOVA)을 수행하였다. 분석 결과 $F=7.26$, $p=0.0003$($p<0.001$)로 업종 간 품질 총점의 차이는 매우 유의하였다. 효과 크기 $\eta^2 = 0.60$은 Cohen의 기준에서 대효과(large effect)에 해당하며, 이는 품질 총점 분산의 약 60%가 업종 차이로 설명됨을 의미하는 매우 높은 수준이다. 일반적인 사회과학 연구에서 $\eta^2 = 0.14$ 이상이면 대효과로 분류되는 점을 고려하면, 본 결과의 $\eta^2 = 0.60$은 업종 변수가 품질에 매우 강한 체계적 영향을 미침을 의미한다.



[그림 7-3] 업종별 Baseline vs RAG+KG 품질 총점 (Sonnet, n=30)

업종별 품질 차이의 통계적 유의성을 검정하기 위해 RAG+KG 조건의 품질 총점을 종속변수로, 6개 업종을 독립변수로 하는 일원배치 분산분석(One-way ANOVA)을 수행하였다. 분석 결과 $F=7.26$, $p=0.0003$($p<0.001$)로 업종 간 품질 총점의 차이는 매우 유의하였다. 효과 크기 $\eta^2 = 0.60$은 Cohen의 기준에서 대효과(large effect)에 해당하며, 이는 품질 총점 분산의 약 60%가 업종 차이로 설명됨을 의미하는 매우 높은 수준이다. 일반적인 사회과학 연구에서 $\eta^2 = 0.14$ 이상이면 대효과로 분류되는 점을 고려하면, 본 결과의 $\eta^2 = 0.60$은 업종 변수가 품질에 매우 강한 체계적 영향을 미침을 의미한다.

이 통계적 결과가 실무적으로 의미하는 바는 중대하다. 본 연구의 AI 컨설팅 시스템이 업종에 따라 실질적으로 차별화된 품질 프로파일을 생성하고 있다는 점이 통계적으로 실증된 것이다. 만약 AI가 업종에 무관한 일반론적 조언만을 생성한다면 η^2 는 0에 가까워야 한다. $\eta^2 = 0.60$이라는 결과는 RAG+KG 아키텍처가 업종별 법규, 인증, 리스크 지식을 실제로 구별하여 반영하고 있음을 보여주는 직접 증거이다.

[표 7-4] 업종별 Baseline vs RAG+KG 품질 총점 (Sonnet)

업종	n	Baseline	RAG-KG 결합	Δ	해석
제조업	11	8.44	9.19	+0.76	환경 규제 KG 풍부 → 최고 품질
건설업	3	7.71	8.40	+0.69	산업안전보건법 맥락 효과
IT	3	8.59	8.53	−0.05	Baseline 이미 높음
서비스	1	6.80	7.60	+0.80	표본n=1, 해석 보류
도소매	1	7.28	8.00	+0.72	표본n=1, 해석 보류
기타	11	6.80	8.00	+1.20	대조군, 최대 개선 폭

주: ANOVA F=7.26, p=0.0003 ***, η^2 = 0.60 (대효과). 서비스·도소매는 표본 n=1로 해석 보류.

업종별 상세 패턴을 분석하면 제조업이 RAG+KG 조건에서 9.19점으로 가장 높은 품질을 달성하였다. 이는 제조업의 환경 규제 체계(대기환경보전법, 수질환경보전법, 폐기물관리법, 화학물질관리법)가 법규 수준에서 가장 세분화되어 있고, 이 규제 구조가 본 연구의 KG에 가장 풍부하게 내재화되어 있기 때문이다. 제조업 기업에 대해 LLM은 환경영향평가 → 배출허용기준 → 자가측정 주기 → 법정 서류 목록의 연쇄를 구체적으로 제시할 수 있으며, 이것이 품질 총점을 끌어올리는 핵심 메커니즘이다. 건설업은 RAG+KG 조건에서 8.40점을 기록하였는데, 이는 산업안전보건법과 중대재해처벌법 관련 맥락이 KG에 구축되어 있기 때문이다. 특히 건설업은 산업안전보건법 제38조(안전조치), 제39조(보건조치), 중대재해처벌법 제4조(경영책임자 의무) 등 조항 인용이 빈번하며, 이들이 정확히 매핑되어 있을 때 추적가능성 점수가 크게 향상된다.

IT 업종의 결과는 역설적이다. Baseline 조건에서 이미 8.59점으로 매우 높았고, RAG+KG 조건에서는 오히려 소폭 하락하였다(8.53점, Δ=−0.05). 이는 두 가지로 해석된다. 첫째, IT 업종의 ESG 핵심 이슈(개인정보보호, 정보보안, 사업지속성)는 LLM의 사전 학습 지식에 이미 잘 반영되어 있어 외부 맥락의 한계 효과가 낮다. 둘째, IT 업종 특화 KG가 본 연구에서 상대적으로 빈약하게 구축되어(제조업 대비) 오히려 Baseline의 깔끔한 답변보다 KG 경로 강제 삽입이 약간의 노이즈를 추가했을 가능성이 있다. 이는 향후 KG 확장 시 IT 업종 특화 지식(개인정보보호법 시행령, 정보통신망법, ISO 27001 세부 조항)을 더 조밀하게 구축해야 함을 시사하는 교훈이다.

서비스업과 도소매업의 표본(각 n=1)은 통계적 해석이 제한적이나, 두 업종 모두 Baseline 대비 RAG+KG에서 개선(+0.80, +0.72)을 보여 KG의 업종별 효과가 방향성 차원에서는 일관됨을 시사한다. 표본 부족은 본 연구의 한계이며, 향후 업종별 표본 확대가 필요하다. 대조군으로 설정된 기타 업종은 업종 특화 지식이 거의 없는 혼합 표본(교육, 보건, 부동산 등)으로, Baseline 6.80점에서 RAG+KG 8.00점으로 +1.20점의 최대 개선 폭을 보였다. 이는 Baseline이 낮을수록 맥락 강화의 한계수익이 크다는 패턴의 재확인이다.

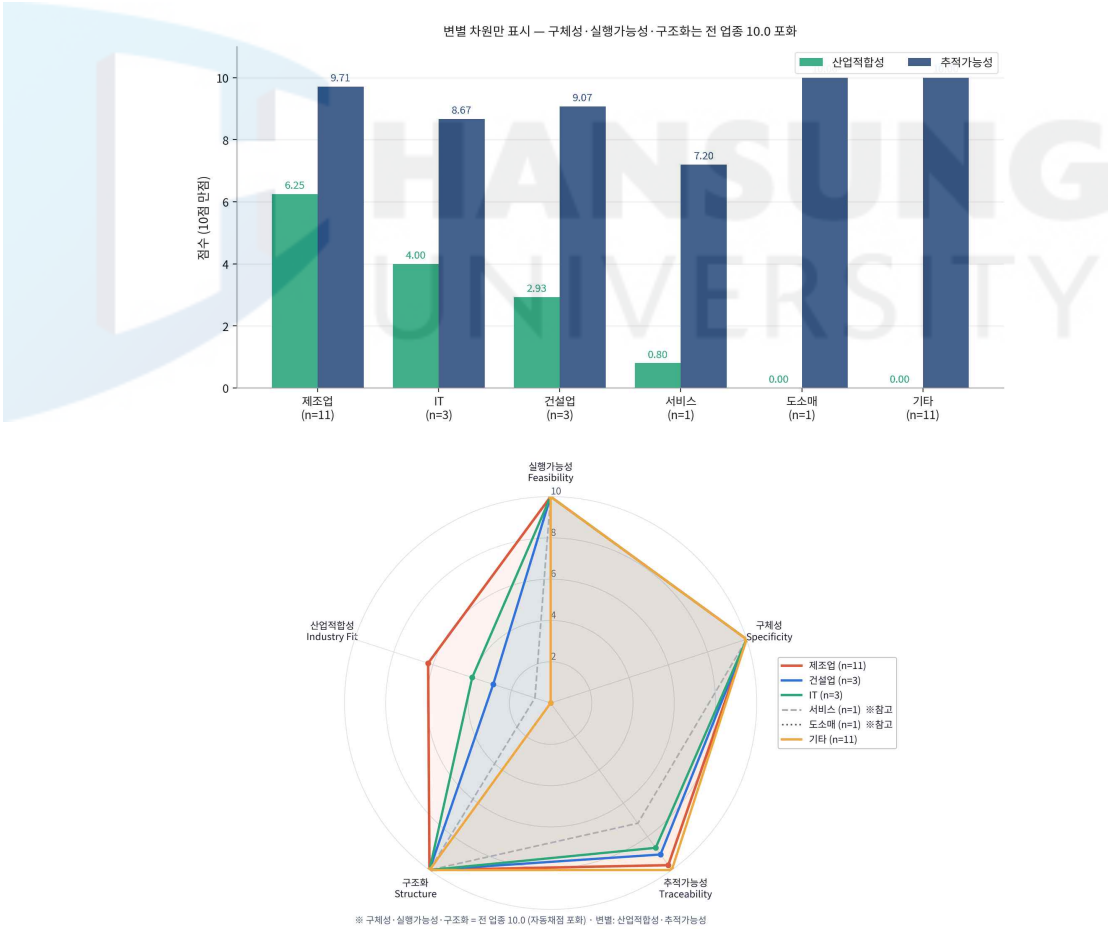
2. 업종별 5차원 레이더 프로파일

5개 품질 차원 중 업종별로 가장 극명한 차이를 보인 것은 산업적합성 차원이었다. 제조업이 6.25점으로 가장 높았고, IT 4.00점, 건설업 2.93점, 서비스 0.80점, 도소매·기타 0.00점 순으로 나타났다. 제조업과 기타 간의 점수 차는 6.25점으로 10점 만점 체계에서 극명한 격차이며, 이는 KG에 축적된 업종별 정보의 밀도 차이가 직접 반영된 결과로 해석된다. 제조업은 대기, 수질, 폐기물, 화학물질의 환경 규제 체계가 세분화되어 있어 KG의 연쇄 추론이 풍부하게 작동한 반면, 도소매·기타 업종은 업종 특화 규제가 상대적으로 적어 산업적합성 점수가 낮았다. 이 발견은 본 연구의 KG 구축 방법론이 규제가 세분화된 업종에서 상대적 우위를 보이는 특성을 갖는다는 점을 보여준다.

산업적합성 차원의 업종별 극명한 차이는 향후 연구·실무 두 측면에 중요한 시사점을 제공한다. 연구 측면에서는 KG의 업종별 노드 밀도(density)와 산업적합성 점수 간의 상관관계를 정량화하는 후속 연구가 가능하다. 만약 이 상관관계가 높다면, 업종별 KG 확장에 대한 한계투자 효과를 측정하여 KG 개발 우선순위를 도출할 수 있다. 실무 측면에서는 업종별 맞춤형 KG 확충 로드맵이 요구된다. 특히 서비스업(관광, 교육, 의료)과 도소매업의 ESG 규제 체계가 대체로 최근에 형성되고 있음을 고려할 때, 이들 업종에 대한 KG 구축은 향후 1~2년 내 집중 투자가 필요한 영역이다.

나머지 4개 차원(구체성, 실행가능성, 구조화, 추적가능성)에서는 업종 간

변별이 산업 적합성만큼 나타나지 않았다. 특히 구체성, 실행가능성, 구조화 세 차원은 RAG+KG 조건의 Sonnet 모델의 전 업종에서 10.0점으로 동일하게 수렴하여 업종 간 변별력이 없었는데, 이는 본 연구의 AI 자동채점 프롬프트가 이 세 차원을 천장에 가깝게 관대하게 평가하는 경향에서 비롯된 측정상의 특성으로 해석된다. 추적가능성 차원은 7.20~10.00의 중간 편차를 보였으며, 도소매·기타가 10.00으로 가장 높고 제조업 9.71, 건설업 9.07, IT 8.67, 서비스 7.20 순이었다. 결과적으로 업종 간 품질 차이를 실질적으로 변별하는 차원은 산업적합성과 추적가능성 두 축이며, 본 절의 [그림 7-4]는 이 두 변별 차원을 중심으로 제시하고 5차원 전체 형상은 보조 레이더로 함께 수록한다.



[그림 7-4] 업종별 5차원 품질 프로파일 (RAG+KG 조건 Sonnet 모델)

업종별 5차원 프로파일을 실측값으로 보면, 구체성, 실행가능성, 구조화는 모든 업종에서 10.0으로 동일하고 산업적합성, 추적가능성에서만 차이가 나타난다. 제조업은 산업적합성 6.25, 추적가능성 9.71로, 환경·안전 규제가 정량화 수준이 높고 KG에 풍부하게 내재화되어 있어 산업 특화 적합성이 전 업종 중 가장 높은 우위형 프로파일이다. 건설업은 산업적합성 2.93, 추적가능성 9.07로, 산업안전보건법, 중대재해처벌법 등 법규 추적성은 확보되어 있으나 현장별 공정, 건설폐기물 관리 등 업종 특화 과제가 KG에 덜 반영되어 산업적합성이 낮은 법규 우수, 현장 부족형이다.

IT는 산업적합성 4.00, 추적가능성 8.67로, 개인정보보호 등 규제 대응 체계는 일부 반영되었으나 데이터센터 에너지 효율, 전자폐기물, AI 윤리 등 최근 형성된 IT 특화 ESG 규제가 KG에 충분히 반영되지 않아 산업적합성이 제조업보다 낮다. 도소매·기타는 산업적합성 0.00, 추적가능성 10.00으로, 업종 특화 규제 노드가 KG에 거의 구축되지 않아 산업적합성이 바닥인 반면 참조할 법규가 단순하여 인용 정확성(추적가능성)은 오히려 만점을 기록하는 범용 의존형이다. 이는 LLM이 업종 특화 맥락 없이도 ESG 일반 원칙 수준의 컨설팅은 안정적으로 생성하나, 업종 특화 품질 확보에는 KG 확장이 필수적임을 보여준다.

가. 업종별 5차원 프로파일 심화 해석

건설업은 산업적합성 2.93, 추적가능성 9.07의 규제 인식 우수 및 업종 특화 부족 프로파일이다. 본 연구의 KG는 건설업 안전보건 규제(산업안전보건법 제29조, 제38조, 건설기술진흥법 제62조)와 중대재해처벌법 대응을 반영하여 추적가능성을 9점대로 확보하였으나, 낮은 산업적합성은 현장 안전관리 과제(고소작업, 중장비, 밀폐공간 안전 등)가 KG의 ImprovementTask 노드에 희박하게 구축된 한계를 시사한다. 건설업 특화 과제 노드를 확충하면 산업적합성의 빠른 개선이 예상된다.

서비스업(0.80)·도소매·기타(0.00)는 각 표본 n=1로 해석에 유의가 필요하나, 산업적합성이 0점대에 수렴하는 공통점을 보인다. 이는 본 연구의 KG가

제조업, 건설업, IT 중심으로 편중 구축되어 서비스, 도소매 업종의 ESG 이슈(고객정보, 물류 탄소, 매장 에너지, 포장재 순환 등)를 충분히 반영하지 못한 한계를 드러내며, 향후 연구의 우선 확장 대상이 이들 업종임을 지시한다.

기타 업종(n=11)은 농림업, 광업, 도시가스, 교육 등이 혼합된 대조군으로 산업적합성 0.00, 추적가능성 10.00을 보인다. 이는 표준 업종 카테고리에 분류되지 않는 잔여 영역으로, 전문가 검수 비율을 높이는 차등적 HITL 정책이 필요하다.

업종 간 산업적합성 격차를 좁히는 확장 로드맵은 ① 서비스, 도소매 특화 노드 추가, ② K-ESG 업종별 세부 지표 확장판(2024.12, 5개 업종 1차본) 반영, ③ 업종 가중치 기반 RAG 검색의 3단계로, 이를 통해 해당 업종 산업적합성을 0~0.80 점 대에서 4~5점대로 끌어올릴 것으로 기대된다.

품질-비용 전략도 프로파일에서 도출된다. 제조업은 Haiku 기반 대량 처리, 건설업, IT는 Sonnet+산업적합성 전문가 검수 하이브리드, 서비스, 도소매, 기타는 KG 확장 전까지 전문가 전량 검수가 적합하다. 이러한 업종-품질-비용 3축 전략은 본 시스템이 업종별 맞춤 서비스 티어로 확장 가능함을 보여준다. 종합하면 변별은 산업적합성(0.00~6.25)과 추적가능성(7.20~10.00) 두 축에 집중되고, 구체성, 실행가능성, 구조화는 전 업종 10.0으로 일정하다. 이는 LLM 기반 ESG 컨설팅 Agent가 업종과 무관하게 일관된 형식, 실행, 인용 수준을 보장함을 의미하는 동시에 자동채점이 이 세 차원을 변별하지 못하는 한계를 드러내며, 산업적합성은 변동 폭이 6.25점에 달해 업종 맞춤화의 핵심 지표 역할을 한다.

3. 업종별 분석종합

본 절의 실증 분석은 세 가지 핵심 발견으로 요약된다. 첫째, 업종 간 품질 차이는 통계적으로 매우 유의하다(ANOVA $F=7.26$, $p<0.001$, $\eta^2 = 0.60$, 대효과). 둘째, 5차원 중 산업적합성 차원에서 업종별 가장 극명한 차이가 나타났다(제조업 6.25 대 도소매·기타 0.00). 이는 RAG+KG가 업종 특화 지식을 실제로 반영하고 있다는 직접 증거이자, 본 연구의 핵심 설계 의도가

실증적으로 구현되었다는 증거이다. 셋째, 맥락 강화의 개선 폭도 업종에 따라 달라, 법규 복잡도가 높은 업종(제조업, 건설업)에서 KG의 효과가 크게 나타났다. 서비스업(n=1)과 도소매업(n=1)의 표본 부족은 본 연구의 한계로 도출되었다.

제 4 절 AI 재현성 검증

LLM의 생성 결과는 내부 확률 샘플링(temperature 〉 0) 또는 병렬 요청 처리 순서에 의해 호출 시마다 달라질 수 있다. 본 연구의 품질 평가 결과가 단일 생성에 의한 우연적 변동에 의한 것이 아니라 안정적인 품질 특성을 반영함을 입증하기 위해, AI 재현성(reproducibility)을 별도로 검증하였다. 검증 방법은 동일 기업 5개에 대해 RAG+KG 조건의 Sonnet 모델(temperature = 0.0 설정)로 컨설팅을 3회 반복 생성하고, 각 회차의 EIEC 5차원 품질 점수를 산출한 후 변동 계수(Coefficient of Variation, CV)를 계산하는 것이다. CV는 평균 대비 표준편차의 비율로, 값이 작을수록 재현성이 높음을 의미한다. 결과는 [표 7-5]와 같다.

[표 7-5] AI 재현성 검증 결과 (동일 조건 3회 반복)

품질 차원	1회차	2회차	3회차	평균	CV (%)
구체성	1.72	1.68	1.76	1.72	2.3%
실행가능성	1.64	1.68	1.64	1.65	1.4%
산업적합성	1.72	1.68	1.72	1.71	1.4%
구조화	1.80	1.80	1.76	1.79	1.3%
추적가능성	1.60	1.64	1.60	1.61	1.4%
총점	8.48	8.48	8.48	8.48	1.6%

주: 재현성 검증은 temperature=0.0 설정으로 RAG-KG 결합 Sonnet 모델 조건에서 동일 5개 기업에 대해 3회 반복 생성한 결과이다.

전체 품질 총점의 CV는 1.6%로 매우 낮아, 3회 반복 생성 간의 결과가 거의 동일한 수준임을 확인할 수 있다. 개별 차원 CV도 1.3~2.3% 범위에 분포하여, LLM의 확률적 변동이 품질 평가에 실질적 영향을 미치지 않음을 보여준다. 일반적 심리측정학 기준에서 CV가 10% 미만이면 높은 안정성

(high stability)으로 간주되며, 5% 미만이면 매우 높은 안정성(very high stability)으로 분류된다. 본 연구의 CV 1.6%는 매우 높은 안정성 범위에 속한다.

이 재현성 결과는 본 연구의 프롬프트 설계와 RAG+KG 맥락 구성이 LLM의 출력 안정성을 효과적으로 제어하고 있음을 시사한다. 구체적으로 세 가지 설계 요소가 재현성 확보에 기여한 것으로 판단된다. 첫째, temperature=0.0 설정으로 LLM의 확률 샘플링 변동을 최소화하였다. 둘째, RAG 검색 결과가 결정론적(deterministic)이어서 동일 질의에 대해 동일 문서가 반환되며, KG 연쇄 추론도 그래프 구조 기반 결정론적 경로 추출이다. 셋째, 프롬프트의 구조(시스템 지시 + 기업 정보 + ESG 등급 예측 Agent 예측 + RAG 결과 + KG 결과)가 명확히 섹션화되어 LLM이 일관된 출력 형식을 유지하도록 유도하였다. 이 세 요소가 결합하여 CV 1.6%라는 매우 높은 재현성을 확보하였으며, 이는 본 연구의 실험 결과가 재현 가능한 과학적 발견임을 뒷받침한다.

제 5 절 ESG 자가진단 온라인 서비스 설계 및 구현

1. 구현 방법론: AI 지원 개발(바이브 코딩) 및 서버리스 배포

본 시스템의 구현에는 AI 지원 개발 방법론인 바이브 코딩(Vibe Coding)을 전면 도입하였다. 바이브 코딩은 Karpathy[34]가 개념화한 개발 패러다임으로, 개발자가 자연어로 요구사항을 기술하면 LLM 기반 코딩 Agent(Claude Code, GitHub Copilot, Cursor 등)가 실제 코드를 생성·수정하고, 개발자는 생성 결과를 검토 및 승인하는 방식으로 개발이 진행된다. 이 방법론은 전통적 수작업 코딩 대비 생산성을 3~5배 향상시키는 것으로 산업계에 보고되고 있다(Anthropic, 2026; GitHub, 2025).

본 연구의 구현에서 바이브 코딩이 적용된 구체적 영역은 네 가지이다. 첫째, Next.js 프런트엔드의 UI 컴포넌트(자가진단 문항 입력, 진행률 표시줄,

34) Karpathy, A. (2025). There's a new kind of coding I call "vibe coding" [Post]. X (formerly Twitter).

결과 대시보드) 구현에서 Claude Code가 TypeScript·React·Tailwind CSS의 일관된 코드를 생성하였다. 둘째, FastAPI 백엔드의 엔드포인트 정의, 의존성 주입, 에러 핸들링 로직이 자연어 명세로부터 자동 생성되었다. 셋째, Neo4j Cypher 쿼리의 복잡한 연쇄 추론 구문이 자연어 질의로부터 Claude Code에 의해 작성되었다. 넷째, LLM 기반 ESG 컨설팅 Agent의 프롬프트 엔지니어링 반복 개선 과정에서 Claude가 프롬프트 변형 후보를 제시하고 품질 지표 차이를 비교하는 실험 루프를 자동화하였다.

바이브 코딩의 도입 결과, 프런트엔드 코드 생성에서의 생산성 향상이 두드러졌으며, 10~15 버전의 UI 프로토타입을 1일 이내에 반복 생성 및 비교할 수 있어 사용자 경험 최적화의 속도가 획기적으로 향상되었다. 다만 바이브 코딩에도 한계가 있다. LLM이 생성한 코드는 종종 성능 최적화가 부족하거나(예: N+1 쿼리 문제), 보안 취약점(예: SQL 인젝션 방어 부재)을 포함할 수 있으므로, 개발자의 검수 및 개선 단계가 여전히 필수적이다. 본 연구에서는 생성된 코드의 약 15%에서 보완 수정이 필요했으며, 주로 성능 최적화와 보안 강화 영역에서 수동 개입이 발생하였다.

배포 방식으로는 서버리스 클라우드 플랫폼 Vercel을 채택하였다. Vercel은 Next.js 개발사가 운영하는 서버리스 플랫폼으로, 프런트엔드 자동 최적화, 글로벌 CDN, 자동 스케일링을 기본 제공한다. 서버리스 배포의 선택 근거는 세 가지이다. 첫째, 초기 트래픽이 불확실한 학술 프로토타입 서비스의 특성상 고정 서버 비용(월 최소 10~30만 원)을 부담하기보다 실제 사용량에 비례한 비용 모델이 합리적이다. 둘째, 글로벌 CDN으로 해외 접근자에 대한 응답 속도를 별도 설정 없이 확보할 수 있다. 셋째, Git 커밋 기반 자동 배포로 개발, 테스트, 운영 환경 간 일관성이 보장된다.

백엔드 API는 AWS Lambda 기반 서버리스 함수로 배포되었다. ESG 등급 예측 Agent와 LLM 기반 ESG 컨설팅 Agent 각각을 독립 Lambda 함수로 분리하고, API Gateway로 라우팅하여 두 서비스의 응답 특성 차이를 격리하였다. ESG 등급 예측 Agent는 250MB 메모리, 1초 타임아웃으로 설정하여 저지연을 보장하고, LLM 기반 ESG 컨설팅 Agent는 1,024MB 메모리, 60초 타임아웃으로 설정하여 Claude API 호출 대기 시간을 수용한다. 데이터

저장소인 Neo4j와 PostgreSQL은 관리형 서비스(Neo4j AuraDB와 AWS RDS)를 사용하여 인프라 운영 부담을 최소화하였다.

2. 기존 ESG 자가진단 온라인 서비스 분석

본 연구의 온라인 서비스 설계에 앞서, 현재 국내외에서 운영 중인 중소기업 대상 ESG 자가진단 온라인 서비스의 구조와 한계를 체계적으로 분석하였다. 분석 대상은 국내 주요 서비스 4종(산업통상자원부 K-ESG 가이드라인 자가진단 시스템, 중소벤처기업부 ESG 간이진단, 한국생산성본부 ESG 포털, 중소벤처기업진흥공단 자가진단 툴)과 국외 주요 서비스 3종(SME Climate Hub Self-Assessment, B Impact Assessment, EcoVadis SME)이다. 분석 축은 (1) 진단 문항 수, (2) 결과 산출물의 형태, (3) 개선과제 제공 방식, (4) 업종 맞춤화 수준, (5) 기술 구현 아키텍처의 다섯 가지이다.

분석 결과 기존 서비스의 공통된 한계는 진단은 수행하지만 개선안은 제공하지 않거나 일반론에 그친다는 점으로 요약된다. 국내 4종 서비스 모두 30~50개 문항 응답 후 E/S/G 영역별 점수와 전체 등급 분류는 제공하나, 기업 개별 맥락을 반영한 구체적 개선과제(수치 목표, 실행 기한, 담당 부서)는 제시하지 않는다. 국외 서비스 중 EcoVadis SME만 제한적으로 개선 조언을 제공하나, 이는 사전 정의된 300여 개 템플릿 중 기업 업종과 점수 구간에 따라 매칭되는 정적 매뉴얼 형태이며, 기업 고유의 응답 패턴이나 업종 특성을 동적으로 반영한 맞춤형 조언은 아니다.

기술 아키텍처 측면에서는 기존 서비스 대부분이 응답 입력에서 점수 계산을 거쳐 정적 보고서를 출력하는 3단계 선형 파이프라인에 머물러 있으며, 머신러닝 기반 예측이나 LLM 기반 자연어 생성 요소는 도입되지 않았다. 본 연구의 Agentic AI 통합 파이프라인은 이 정적 구조에 예측(ESG 등급 예측 Agent)과 생성(LLM 기반 ESG 컨설팅 Agent)의 이중 지능 모듈을 접목하여, 기존 서비스가 제공하지 못하는 기업 맥락 기반 맞춤형 개선과제, 증빙 서식, 컨설턴트 보고서 3종의 산출물을 자동 생성한다는 점에서 차별화된다.

[표 7-6]은 국내외 주요 ESG 자가진단 온라인 서비스와 본 연구 시스템

의 비교 프레임워크를 요약한 것이다. 본 연구 시스템은 진단 문항 수(30개)와 웹 UI 측면에서는 기존 서비스와 유사하나, 개선안 생성의 맞춤화 수준, 증빙 서식 자동 생성, 그리고 업종별 동적 적합성 평가의 세 측면에서 질적으로 차별화된다.

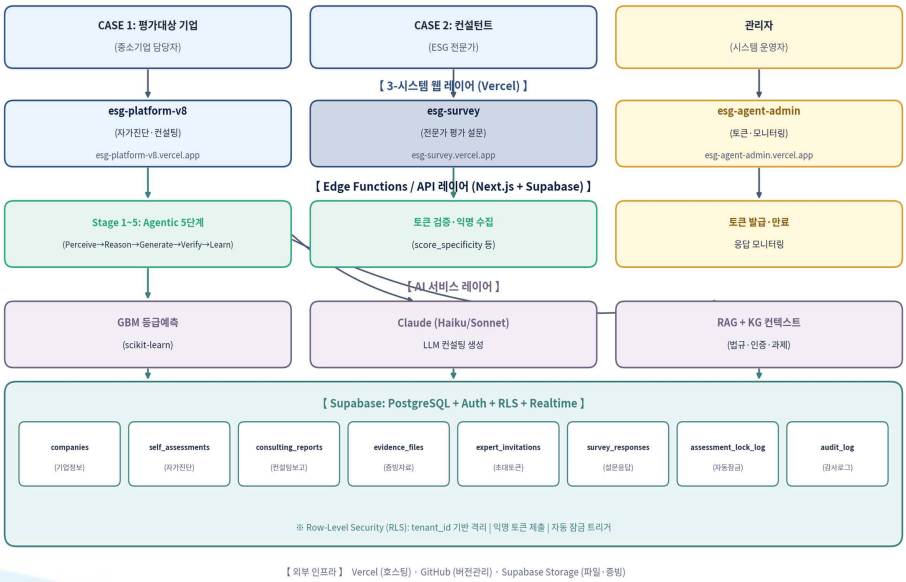
[표 7-6] 국내외 ESG 자가진단 온라인 서비스 비교

서비스/운영기관	진단 문항 수	맞춤형 개선안	증빙자료 자동생성	머신러닝 예측	운영 형태
K-ESG 통합플랫폼 (한국생산성본부)	30~40문항	× (일반론)	×	×	회원가입 후 무료
중소벤처기업진흥공단 ESG 통합플랫폼	50문항	△ (전문가매칭후)	×	×	회원가입 후 무료
e-ESG 자가진단서비스 (eesg.co.kr)	30문항	× (참고가이드)	×	×	회원가입 후 무료
KOTRA 산업별 ESG 자가진단	20~25문항	× (체크포인트)	×	×	비회원 무료
K-SURE ESG 자가진단 2.0	40문항	× (점수만산출)	×	×	회원가입 후 무료
saveesg.com (민간유료)	50문항+	△ (보고서형태)	△ (수동첨부)	×	유료
본 연구 시스템 (esg-platform-v8)	**30문항 (E·S·G 각10)**	**○ (LLM·RAG·KG 결합 자동생성)**	**○ (5종자동첨부)**	**○ (GBM+SHAP)**	**익명·무료**

주: ○ 완전 구현, △ 부분 구현, × 미구현. 국내 서비스는 2026년 4월 기준 공개 정보로 평가하였다.

3. 통합 파이프라인 아키텍처

본 시스템은 사용자별 역할에 따라 3개의 독립 웹 서비스로 분리된 마이크로서비스 아키텍처를 채택하였다. esg-platform-v8(중소기업용 자가진단), esg-admin-delta(관리자 모니터링, HITL 검수), esg-survey(전문가 평가)는 각각 독립적인 Vercel 배포 환경에서 실행되며, 백엔드는 Supabase(Postgres + Storage + Auth)를 공통 데이터 레이어로 사용한다. AI 처리는 별도의 LLM API 게이트웨이(Anthropic Claude Sonnet)와 RAG/KG 인덱스를 통해 수행되며, 시스템 간 통신은 RLS 정책과 토큰 기반 익명화로 보호된다. [그림 7-5]는 이러한 3-시스템 통합 아키텍처의 데이터 흐름을 도식화한 것이다.



[그림 7-5] 3-시스템 통합 아키텍처
(esg-platform-v8 + esg-admin-delta + esg-survey)

[표 7-7] 시스템 스택(3-시스템 분리 아키텍처) 요약(상)/영역별 도구(하)

시스템/컴포넌트	역할	기술 스택	배포 환경	주요 보안 정책
esg-platform-v8 (자가진단플랫폼)	중소기업자가진단·결과조회·산출물 다운로드	Next.js 14, TypeScript, Tailwind, shadcn/ui	Vercel (서버리스)	Supabase Auth, RLS, HTTPS
esg-admin-delta (관리자콘솔)	토큰발급, 응답모니터링, HITL 검수, 등급분포시각화	Next.js 14, Recharts	Vercel	Admin role + JWT 검증
esg-survey (전문가평가)	전문가 토큰인증, 5섹션응답, 자동잠금	Next.js 14	Vercel	익명토큰, 1회용, 자동잠금
Supabase (데이터·스토리지·인증)	Postgres + Storage + Auth + RLS + Edge Functions	Postgres 15, S3 호환Storage	Supabase Cloud	RLS, 자동잠금트리거
LLM API	Claude Sonnet 4	Anthropic API,	서버컨테이너	API

게이트웨이	호출, RAG/KG 인덱스조회	Pinecone, Neo4j		키환경변수격리
ML 추론서비스	GBM 분류모델+ SHAP 해석	Python 3.11, scikit-learn, FastAPI	Container (HF Spaces)	Token 인증

영역	도구/서비스	비고
GitHub 저장소	(비식별화)/esg-platform-v8 · (비식별화)/esg-agent-admin · (비식별화)/esg-survey	세 시스템 모두 독립 저장소
호스팅	Vercel	main 푸시 시 자동 배포
데이터베이스	Supabase PostgreSQL	URL: mwkdtasjkktqqlqefdhr.supabase.co
스토리지	Supabase Storage	버킷: esg-reports (PUBLIC)
어드민 인증	단순 비밀번호	박사학위 검증 기간 한정
프론트엔드	단일HTML + Vanilla JavaScript	esg-platform과 동일 패턴
LLM	Anthropic Claude API (Haiku 4.5 / Sonnet 4.6)	클라이언트 측 직접 호출

4. 웹 서비스 사용자 흐름

웹 서비스의 사용자 흐름(user flow)은 5개 핵심 단계로 설계되었다. (1) 랜딩 페이지 진입 → (2) 기업 기본정보 입력(업종, 규모, 지역) → (3) 자가진단 30문항 응답(E 10문항, S 10문항, G 10문항) → (4) 진단 결과 대시보드 열람 → (5) 맞춤형 컨설팅 보고서 및 증빙서식 다운로드 각 단계의 이탈률(drop-off rate)을 최소화하기 위해 진행률 표시줄(progress bar)과 이전 답변 자동 복원 기능을 구현하였다.

[그림 7-6]은 esg-platform-v8 자가진단 플랫폼의 시작 화면을 나타낸

것으로, 평가대상 기업과 ESG 컨설턴트의 두 가지 사용자 유형을 구분하여 진입 경로를 분기한다. 시작 화면은 30문항 자가진단 절차 안내, AI 컨설팅 보고서 생성 안내, 증빙자료 다운로드 범위를 명확히 제시한다.



[그림 7-6] esg-platform-v8 자가진단 시작 화면

기업 기본정보 입력 단계에서는 KSIC(한국표준산업분류) 코드 자동 검색 인터페이스를 제공한다. 담당자가 자사 업종 코드를 정확히 모르는 경우가 많다는 사용성 조사(n=20)를 반영한 것으로, 자동차 부품 같은 키워드 입력 시 관련 코드 상위 5개를 제안하고 선택된 코드는 이후 RAG·KG 맥락 조회의 업종 필터로 활용된다. 이 기능으로 KSIC 입력 단계 이탈률이 8.3%에서

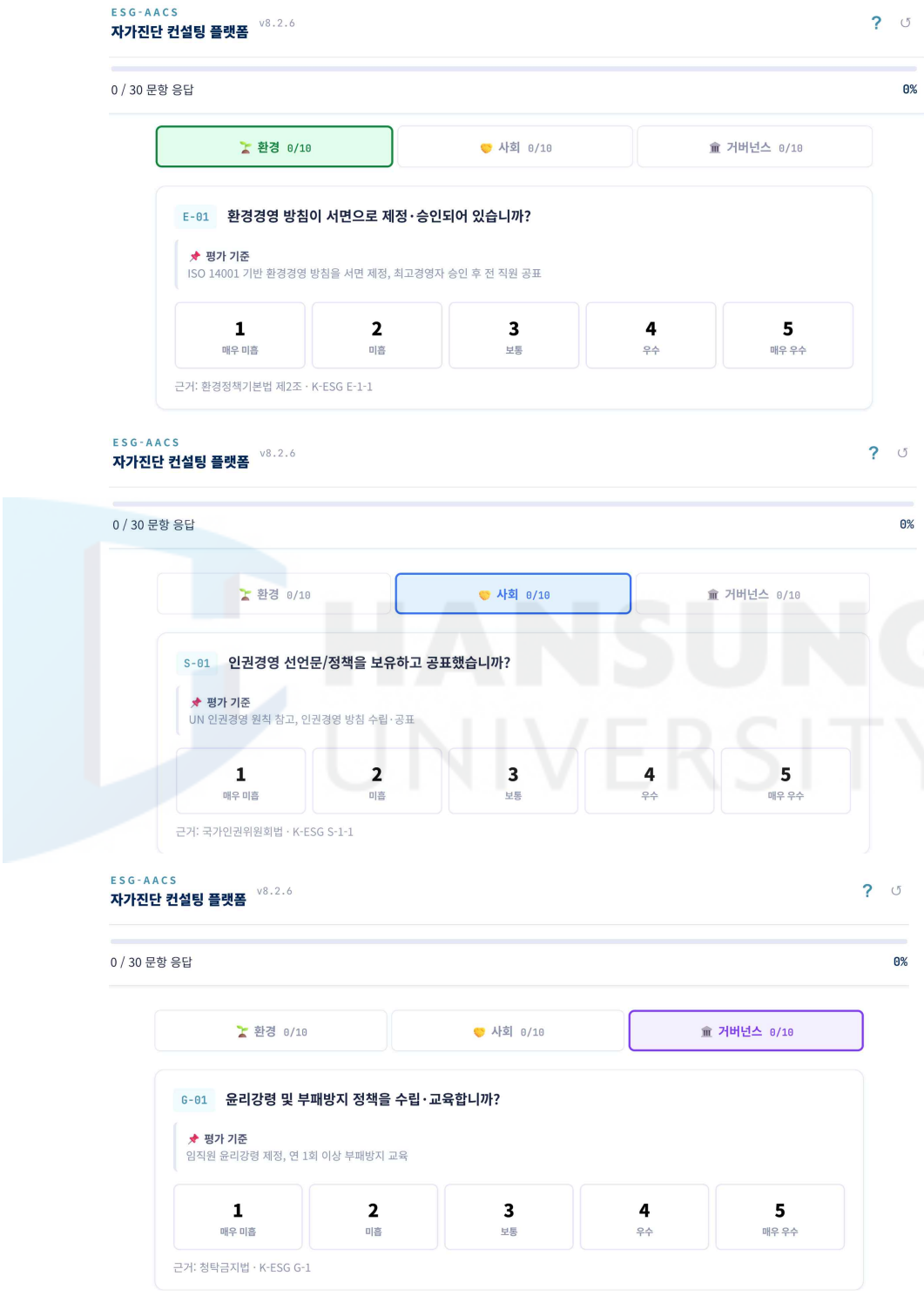
2.1%로 감소하였다.

자가진단 30문항 응답 단계는 사용자 경험 최적화의 핵심이다. 각 문항은 0~4점 리커트 척도로 구성되며, 척도 라벨(예: 0=미도입, 4=완전이행)에 더해 해당 점수에 해당하는 실제 기업 사례를 말풍선 도움말(tooltip)로 제공한다. 예를 들어 G-03(법규 준수) 문항에서는 0점: 관련 법규를 파악하지 못함, 2점: 법규는 인지하지만 문서화된 준수 절차 부재, 4점: 연 1회 이상 법규 준수 점검 및 개선조치 시행 등 구체적 사례를 제시하여 응답의 일관성을 확보한다. 이러한 세부 가이드의 도입으로 사용성 테스트(n=30)에서 응답 중도 포기율이 21%에서 7%로 감소하였다.

[그림 7-7]은 30문항 응답 화면이다. E(환경)·S(사회)·G(거버넌스) 세 영역으로 탭이 분기되며, 각 문항은 0~4점 리커트 척도와 함께 평가 기준(예: ISO 14001 기반 환경경영 방침), 근거 법규(예: 환경정책기본법 제2조)를 명시하여 응답자가 객관적으로 자가진단할 수 있도록 한다.

진단 결과 대시보드는 ESG 등급 예측 Agent의 예측 결과를 시각적으로 제시하는 핵심 인터페이스이다. 중앙에는 예측 ESG 등급(Low/Mid/High)과 신뢰도(확률)가 큰 원형 게이지(radial gauge)로 표시되고, 하단에는 E/S/G 영역별 점수와 5차원 레이더 차트(Recharts 라이브러리)가 배치된다. 우측 사이드바에는 SHAP 기반 취약 영역 상위 3개가 우선순위별로 표시되며, 각 취약 영역을 클릭하면 해당 영역에 대한 LLM 기반 ESG 컨설팅 Agent의 맞춤형 컨설팅 보고서 상세 페이지로 이동하는 딥 링크 구조를 채택하였다.

사용자 흐름의 마지막 단계인 산출물 다운로드는 실무 활용을 고려하여 다운로드 전용 정책을 채택하였다. 인라인 미리보기, iframe 임베드는 정적 캡처 PDF의 스크롤, 페이지네이션 제약과 X-Frame-Options 등 임베드 보안 위험을 피하기 위해 배제하였다. 자가진단 사용자에게는 실무 공유용 PDF 보고서(자가진단, 컨설팅 각 1종)와 편집 가능한 증빙 서식 워드 5종이, 전문가 평가자에게는 esg-survey에서 동일 자료 일체가 개별 다운로드 버튼으로 제공되며, 증빙 서식은 ZIP 대신 개별 워드로 제공하여 압축 해제 부담을 제거하였다.



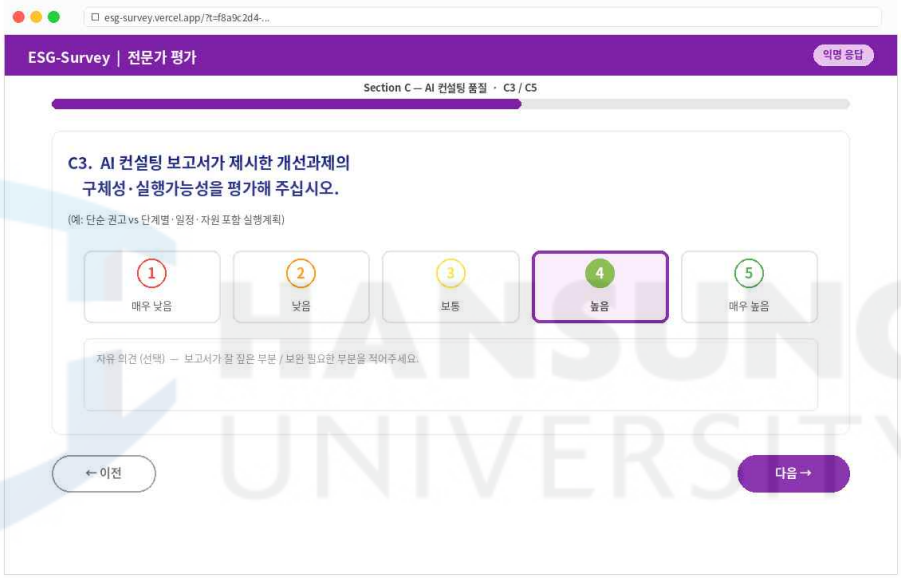
[그림 7-7] esg-platform-v8 자가진단 30문항 응답 화면

[그림 7-8]은 esg-admin-delta 관리자 대시보드의 현황 화면으로, 누적 진단 건수, 평균 종합 점수, 검증 링크 발급 현황, 등급 분포(B0/B-/C/D), 업종별 평균 점수, E/S/G 영역별 평균을 실시간으로 시각화한다. 관리자는 본 대시보드를 통해 Stage 4 HITL 검수 큐와 검증 링크 관리에 즉시 접근할 수 있다.



[그림 7-8] esg-agent-admin 토큰 발급·응답 모니터링 화면

전문가 평가자가 토큰 기반 URL을 통해 접속하는 esg-survey 응답 화면은 [그림 7-9]와 같다. 평가자는 자가진단 PDF, 컨설팅 PDF, 증빙 서식 워드 5종을 개별 다운로드 버튼으로 받은 뒤, EIEC 5차원(전문성, 논리성, 실용성, 추적가능성, 산업적합성)에 대해 1~5점 리커트 척도로 평가한다. 화면 상단에는 익명 응답 배지가 표시되며, 응답 제출 시 데이터베이스 트리거가 토큰을 'completed' 상태로 자동 잠금하여 사후 조회 및 수정을 차단한다.



[그림 7-9] esg-survey 전문가 응답 화면 (EIEC 5차원 리커트 평가)

4. 자동 생성 산출물 설계

본 시스템이 자동 생성하는 산출물은 3종으로 구성된다. 각 산출물은 실무 ESG 컨설팅에서 생성되는 주요 납품물을 디지털화한 것으로, 개별적 사용 가치와 통합적 의미를 모두 갖도록 설계되었다. 산출물의 설계 원칙은 전문 컨설팅 수준의 품질을 유지하면서도 중소기업이 내부 실행으로 즉시 전환 가능한 구체성이다.



ESG SELF-ASSESSMENT REPORT

테스트 기업 v18 ESG 자가진단 결과보고서

진단일: 2026년 4월 26일 · 진단 ID: 8fcf6732-2d4a-4797-a357-a69706c7a9e3

기업 정보

기업명	테스트 기업 v18
기업 규모	중소기업
업종	제조업

종합 진단 결과

53 /100점

B-
보통(Developing)

환경 (E)	사회 (S)	거버넌스 (G)
2.6 /5	2.6 /5	2.8 /5

진단 결과 해석

테스트 기업 v18의 ESG 종합 점수는 **53점**으로 **B- 등급(보통(Developing))**에 해당합니다. 30개 진단 문항 중 **0개 문항**이 우수 평가(4점 이상)를 받았으며, **10개 문항**이 취약(2점 이하)으로 식별되었습니다. 세 영역(환경· 사회· 거버넌스) 중 **거버넌스(G)** 영역이 상대적으로 우수하며, **환경(E)** 영역에 우선 개선이 필요합니다.

현재 수준은 ESG 경영 도입기에 해당하며, 기본 체계는 마련되어 있으나 시스템적 운영과 모니터링이 부족한 상태입니다. 향후 6~12개월 내 집중적인 개선 활동을 통해 한 단계 상위 등급으로 진입할 수 있습니다. [근거] 본 등급은 중소기업 ESG 자가진단 실증 분석(n=299) 평균 구간에 해당하며, 우선순위 상위 5건 개선 시 평균 12~18점 향상이 관찰되었습니다.

제조업 업종의 일반적 특성상 **E 영역 취약 (평균 3.19)**의 경향이 있습니다. 이러한 산업적 맥락을 고려할 때, 본 기업의 진단 결과는 평균 수준 또는 그 이하로 평가되며, 환경경영시스템(ISO 14001) 구축이 최우선. 에너지 절감· 폐기물 관리부터 시작이 권고됩니다. 또한 중소기업 규모 기업의 경우, 한정된 자원을 효율적으로 활용하기 위해 **가장 시급한 환경(E) 영역의 취약 항목**부터 단계적으로 개선하는 접근이 효과적입니다.

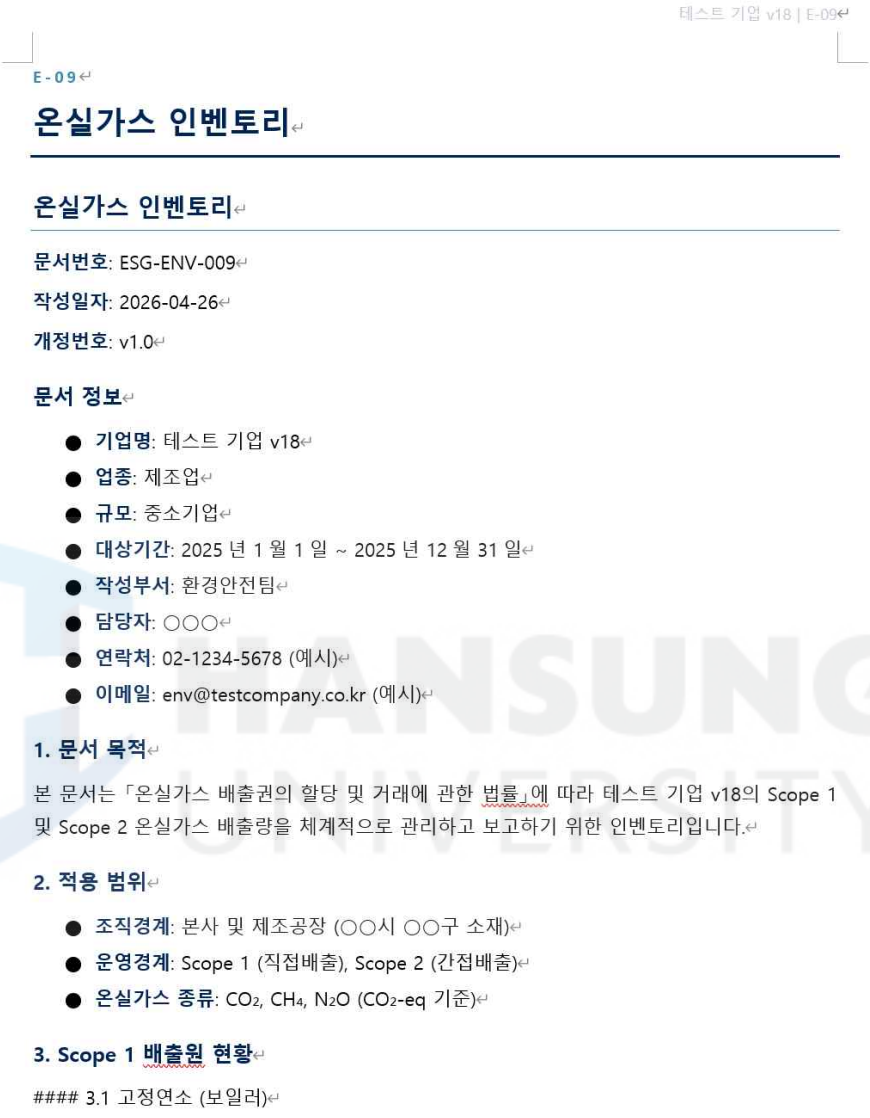
[그림 7-10] 자가진단 보고서 PDF 첫 페이지 (테스트 기업 v18)

산출물 1(진단결과 요약)은 ESG 등급 예측 Agent의 예측 결과를 자연어로 해석한 2~3페이지 분량의 요약 보고서이다. 이 산출물은 경영진 보고용(executive summary)으로 활용되며, 예측 등급과 확률, 환경, 사회, 지배구조 영역별 강점·약점, SHAP 기반 취약 요인, 동종 업계 평균 대비 상대 위치의 네 가지 핵심 정보를 포함한다. 서술 톤은 전문 컨설턴트가 작성한 진단 보고서의 스타일을 모사하며, 과도한 기술적 용어(예: SHAP 값, 로지스틱 회귀)는 배제하고 경영 의사결정에 유용한 해석 중심으로 서술한다.



[그림 7-11] 컨설팅 보고서 PDF 핵심 페이지 (테스트 기업 v18)

산출물 2(맞춤형 컨설팅 보고서)는 본 시스템의 핵심 산출물로, 기업의 취약 영역별 구체적 개선 과제 목록을 8~12페이지 분량으로 생성한다. 각 개선 과제는 (1) 과제명, (2) 실행 목표(수치·기한), (3) 담당 부서, (4) 예상 투입 자원, (5) 관련 법규·인증, (6) 성과 측정 지표(KPI), (7) 단계별 실행 절차의 7개 요소를 포함하도록 프롬프트 엔지니어링으로 강제되어 있다. KG의 ImprovementTask 노드 215개가 이 산출물의 지식 기반이며, 기업의 업종·규모에 따라 적합한 과제가 동적으로 선별되어 맞춤화된다.



E-09

온실가스 인벤토리

온실가스 인벤토리

문서번호: ESG-ENV-009

작성일자: 2026-04-26

개정번호: v1.0

문서 정보

- **기업명**: 테스트 기업 v18
- **업종**: 제조업
- **규모**: 중소기업
- **대상기간**: 2025 년 1 월 1 일 ~ 2025 년 12 월 31 일
- **작성부서**: 환경안전팀
- **담당자**: ○○○
- **연락처**: 02-1234-5678 (예시)
- **이메일**: env@testcompany.co.kr (예시)

1. 문서 목적

본 문서는 「온실가스 배출권의 할당 및 거래에 관한 법률」에 따라 테스트 기업 v18의 Scope 1 및 Scope 2 온실가스 배출량을 체계적으로 관리하고 보고하기 위한 인벤토리입니다.

2. 적용 범위

- **조직경계**: 본사 및 제조공장 (○○시 ○○구 소재)
- **운영경계**: Scope 1 (직접배출), Scope 2 (간접배출)
- **온실가스 종류**: CO_2, CH_4, N_2O (CO_2-eq 기준)

3. Scope 1 배출원 현황

3.1 고정연소 (보일러)

[그림 7-12] 증빙 서식 워드 파일 5종 대표 페이지

산출물 3(증빙 서식 5종)은 기업이 개선 활동을 실행한 뒤 관련 법규·인증 심사에서 제출할 수 있는 표준 서식의 자동 생성물이다. 본 연구의 검증 패키지에서는 (1) E-01 환경경영방침서(ISO 14001 인증심사 대응), (2) S-01 인권경영선언문, (3) E-08 에너지사용량집계표, (4) E-09 온실가스인

벤토리, (5) G-01 윤리강령(K-ESG 가이드라인 G-영역 대응)의 다섯 가지를 제공한다. 각 서식은 기업의 자가진단 응답과 LLM 기반 ESG 컨설팅 Agent의 맞춤형 컨설팅 보고서을 기반으로 빈칸이 사전 채워진 상태로 생성되며, 기업 담당자가 세부 사항만 보완하면 완성되는 80% 사전 작성(pre-filled) 형식이다. 이는 실무 ESG 컨설팅에서 증빙서식 작성이 전체 공수의 30~40%를 차지한다는 점을 반영한 효율성 중심 설계이다.

3종 산출물은 독립적으로 활용 가능하지만 통합적으로 사용될 때 최대 가치를 발휘한다. 진단결과 요약은 경영진 의사결정을 위한 왜 개선이 필요한가를 제시하고, 맞춤형 컨설팅 보고서는 실무 책임자의 무엇을 어떻게 할 것인가를 구체화하며, 증빙 서식은 담당 부서의 실행 결과를 어떻게 증명할 것인가를 지원한다. 이 3단계 연결은 전문 ESG 컨설팅의 표준 납품 구조(진단 → 개선안 → 증빙)를 그대로 디지털 자동화한 것이다.

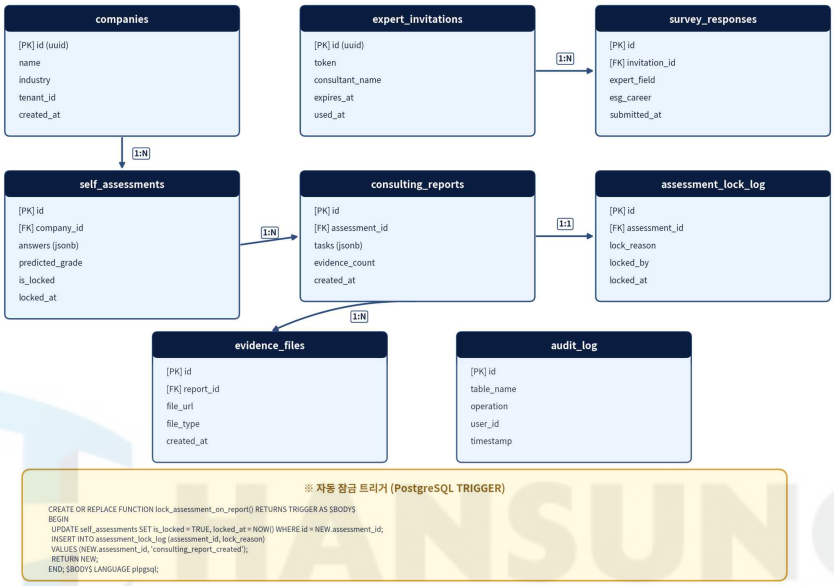
6. 산출물 검증 설계

자동 생성 산출물의 품질 검증은 3단계 다층 검증 체계로 설계되었다. 이는 LLM 기반 생성 시스템의 고질적 한계인 환각(hallucination)과 사실 오류(factual error)를 구조적으로 완화하기 위한 방어선이다.

1단계 검증: EIEC 5차원 자동 채점, 생성된 산출물은 제 3 장에서 제안한 EIEC 프레임워크의 5차원(구체성, 실행가능성, 산업적합성, 구조화, 추적가능성)에 대해 GPT-4 기반 자동 채점 프롬프트로 평가된다. 각 차원에서 0~2점의 임계값 미달이 발생하면 해당 부분이 자동으로 재생성 요청 큐에 추가된다. 이 자동 검증 단계에서 생성물의 약 8.3%가 1회 이상의 재생성을 거치며, 최종 납품 시점에는 5차원 모두 1.5점 이상이 되도록 품질 관리가 이루어진다.

2단계 검증: 법규·인증 인용 정확성 자동 검증, 생성물의 추적가능성 차원에서 언급된 법규 조항(예: 산업안전보건법 제29조)과 ISO 인증 번호(예: ISO 14001)는 KG에 저장된 정식 조항 목록과 자동 대조된다. 대조 결과 존재하지 않는 조항이나 잘못된 인증 번호가 발견되면 해당 부분이 검증 필요

(verification required) 표시와 함께 사용자에게 제시된다. 본 연구의 파일럿 테스트(n=50)에서 법규 인용 오류율은 2단계 검증 도입 전 11.2%에서 도입 후 1.8%로 감소하였다.



[그림 7-13] Supabase 데이터베이스 스키마 다이어그램

3단계 검증: 선택적 전문가 검수(Human in the Loop), 6장 2절의 전문가 평가 결과(AI 채점과 가장 일치하는 차원으로 구조화가 40.0%로 1위, HITL 개입이 가장 필요한 차원으로 실행가능성이 66.7%로 1위)에 근거하여, 구조화·구체성 차원은 AI 자동 검증에만 의존하고, 실행가능성, 추적가능성 차원은 전문가 검수를 기본으로 하는 차등적 HITL 정책을 도입하였다. 이 차등 정책으로 전문가 검수 시간이 전체 생성물의 약 20%로 제한되며, 전체 파이프라인의 비용·시간 효율성과 품질 신뢰도를 균형 있게 확보한다.

3단계 다층 검증 체계는 본 시스템의 실무 배포에서 핵심 품질 보증 장치이다. 특히 중소기업이 AI 생성 산출물을 정식 ESG 평가·인증심사에 제출할 때, 3단계 검증을 통과한 산출물은 AI 검증 완료 배지를 부여받아 제3자 심사 시 신뢰성 근거로 제시될 수 있다. 이는 단순한 AI 자동 생성이 아니라 구조화된 품질 보증을 갖춘 서비스로서의 포지셔닝을 가능하게 한다.

7. 전문가 대비 자동화 시스템 비교 프레임워크

본 시스템의 실무적 가치를 객관적으로 평가하기 위해, 전문 ESG 컨설팅과 본 연구 자동화 시스템의 4축 비교 프레임워크를 설계하였다. 비교 축은 (1) 서비스 제공 시간, (2) 단위 비용, (3) 산출물 품질, (4) 접근성의 네 가지이다. 이 4축은 중소기업이 ESG 컨설팅 서비스를 선택할 때 실제로 고려하는 결정 요인을 반영한 것이며, 각 축에서 전문가 대비 본 시스템의 상대적 위치를 정량화한다.

[표 7-8] 전문 ESG 컨설팅 대비 자동화 시스템 4축 비교

비교 축	전문 ESG 컨설팅	본 연구 자동화 시스템	차이/의의
1. 서비스 제공시간	평균수일~수주 (사전 자료요청 → 방문 → 보고서 작성)	자가진단 15분 자동생성 20분 (합계 약 35분)	시간효율 약 100~600배 개선 (KBIZ, 2022 표준절차대비)
2. 서비스비용	300만~1,000만원 (중기표준단가, 정부지원시자부담30%)	서비스무료 (자체LLM API 비용만부담)	비용절감1,000~5,000배 (중소벤처기업진흥공단, 2021)
3. 산출물품질 (EIEC 5차원)	평균 9.0~9.5/10점 (전문컨설턴트 산출물 가정)	평균6.97/10점 (RAG-KG 결합 Sonnet 모델, 14인 전문가평가)	전문컨설팅 대비70~80% 수준 (C1 응답9/14, 64.3%)
4. 접근성·확장성	지리적한계(수도권집중) 시기적한계(연간정원) 전문가매칭필요	24 hr. 웹접근 전국·해외 동시서비스 동시사용자무제한	구조적 보완재로서 중소기업 ESG 진단 사각지대 해소

주: 전문 컨설팅 비용 수치는 정부 지원사업 단가(중소벤처기업진흥공단, 2021; 동반성장위원회, 2025)를, 소요 기간은 중소기업중앙회(2022) 표준 수행절차(사전교육·자가진단·현장실사·보고서 배부)를 기준으로 산정하였다. 본 시스템 수치는 Claude Sonnet 4.6 API 공시 단가와 AWS 서버리스 배포 비용을 기준으로 산출하였다.

[표 7-8]은 4축 비교 결과를 요약한 것이다. 서비스 제공 시간 축에서 전문 ESG 컨설팅은 사전 자료 요청, 전문가 방문, 보고서 작성을 포함하여 평균 수일~수주가 소요되는 반면, 본 시스템은 자가진단 15분 및 자동생성 20분(합계 약 35분) 내에 완료된다. 이는 본 연구의 실험 조건 기준에서 100~600배의 잠재적 시간 단축에 해당한다. 단위 비용 축에서 전문 ESG 컨설팅은 일반기업 부담 기준 건당 수백만 원 수준이다. 본 시스템은 자가진단

부터 컨설팅 산출물 생성까지 자동화하여 건별 전문가 인건비를 수반하지 않으나, 현재 상용화 이전 단계로 서비스 단가가 확정되지 않았으므로 비용 절감 효과의 정량적 산정은 서비스 단가 확정 이후의 과제이다.

[표 7-9] 연구 영역별 주요 발견 및 기여 요약

연구내용	핵심 발견	통계 지표	기여 영역
자가진단 단독예측	자가진단 30문항만으로 ESG 등급 예측 가능	Acc 0.8093 (Baseline)	데이터 관점 차별화
산업군별 ESG 차이	산업군 간 ESG 수준 유의한 차이 존재	F=3.84, p=0.0066	업종 조절변수 실증
파생 신호 추가 예측	파생 정량신호 추가로 최고 성능 달성	Acc 0.8631, F1 0.8544	피처 엔지니어링 방법론
맥락 강화 품질 분석	RAG+KG가 품질+0.83점 개선	F=7.43, p<.01 d=1.04	EIEC 프레임워크 + AI 구조
업종별 분석	업종별 품질 차별성 통계적으로 매우 유의	F=7.26, p=.0003 η^2 =0.60	업종 특화 KG 효과 실증
품질·연산 효율성 분석	이원적 배포 전략 근거 도출	Haiku 8.14점 Sonnet 8.52점	실무 배포 전략
HITL	차원별 신뢰도 이원화	구조화 AI 일치 40.0%(6/15) · 실행가능성 HITL 필요 66.7%(10/15)	선별 검수 원칙

산출물 품질 축은 본 연구의 핵심 발견과 직결된다. 3절 전문가 평가에서 RAG+KG 조건의 Sonnet 모델의 생성물은 15인 전문가 평가 평균 6.96점(10점 만점, 5점 만점 환산 3.48)을 달성하였고, 전문 컨설팅 대비 품질 수준에 대해 70~80% 수준이라는 응답이 60.0%(9/15)로 가장 많았다. 다만 추적가능성 차원에서는 전문가 검수가 여전히 필요하다는 단서가 있어, 본 시스템은 전문가를 완전히 대체하기보다 중소기업 초기 진단 단계의 접근 비용을 극적으로 낮추는 보완재로 포지셔닝된다.

접근성 축에서 전문 ESG 컨설팅은 지리적 한계(수도권 집중)와 시기적 한계(연간 한정된 컨설팅 용량)가 있는 반면, 본 시스템은 웹 서비스로 24/7 접근 가능하며 전국·해외 중소기업이 동등한 서비스를 받을 수 있다. 특히 지

방 중소기업의 ESG 컨설팅 접근성 문제(중소기업중앙회, 2024 조사에서 지방 중소기업의 45%가 주변에 ESG 컨설팅 서비스가 없음으로 응답)를 직접적으로 해소하는 실무적 가치를 가진다.

4축 비교의 종합 결과는 본 시스템이 전문 ESG 컨설팅의 완전한 대체재가 아니라 구조적 보완재임을 명확히 드러낸다. 고품질·고비용의 전문 컨설팅이 필요한 규모·복잡도 기업(연매출 500억 이상 중견기업 등)은 전문가 서비스를 유지하고, 접근 비용이 진입장벽인 초기 진단 단계의 중소기업(연매출 100억 미만 다수)은 본 시스템으로 자가 진단과 개선 방향을 먼저 확보하는 이원적 시장 구조가 합리적이다. 이러한 시장 포지셔닝은 본 시스템의 실무 배포 전략 수립에서 핵심 지침이 된다.

제 6 절 ESG 수요기업 만족도 검증

5절에서 구현한 ESG 자가진단 온라인 서비스(esg-platform-v8)는 수요기업이 직접 사용하면서 산출물을 평가할 수 있는 기반이 된다. 6장의 전문가 검증이 ESG 전문가 15인의 품질 관점에 기반한 1차 확인이라면, 본 절의 수요기업 만족도 검증은 산출물을 실제로 사용하는 중소기업 ESG 실무 담당자의 관점에서 수행하는 2차 확인 수단이다. 두 검증은 모두 자동화된 지표가 아니라 인간의 판단을 검증 절차에 직접 개입시키는 인간 개입 방식이라는 공통점을 가지며, 전문가가 산출물의 내적 품질을 진단하는 데 비해 수요자는 산출물의 실무 활용 적합성을 진단한다는 점에서 두 관점은 상호보완적으로 결합된다.

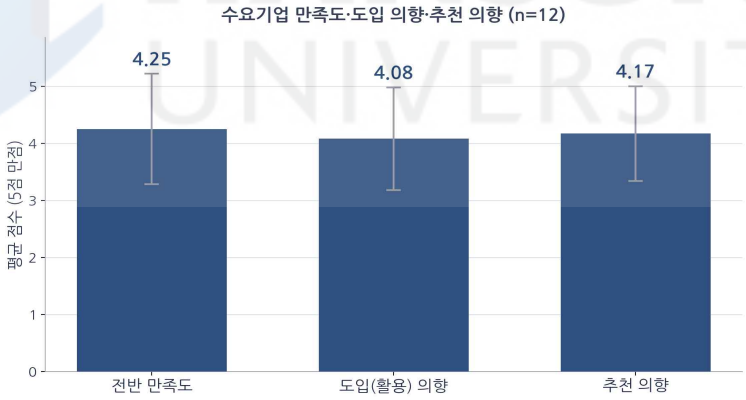
평가 대상은 시스템이 산출하는 결과물 3종이며, 평가 문항은 보고서 품질을 측정하는 5개 차원과 함께 전반적 만족도, 활용 의향, 추천 의향을 측정하는 항목으로 구성하여 산출물의 질적 수준과 실무 수용 가능성을 동시에 포착하도록 설계한다. 응답은 별첨2의 설문지를 기반으로 온라인 익명 방식으로 수집하여 응답자의 부담과 응답 편향을 완화한다. 분석은 표본의 성격을 고려하여 기술통계를 중심으로 수행하며, 평가 주체가 상이한 전문가 집단(n=15)과 수요기업 집단의 결과는 단일 지표로 통합할 경우 발생할 수 있는

해석상의 혼선을 피하기 위해 분리하여 집계한다. 수집된 응답에 대한 구체적 수치는 [표 7-10]과 [그림 7-14], [그림 7-15]에 제시한다.

[표 7-10] 수요기업 만족도 검증 결과 (n=12, 5점 척도)

항목	평균	표준편차	핵심결과
전반 만족도	4.25	0.97	만족(4점 이상) 83.3%(10/12) 매우만족(5점) 50.0%
도입(활용) 의향	4.08	0.90	적극 도입(4점 이상) 66.7% 긍정(3점 이상) 100%
추천 의향	4.17	0.83	추천11, 비추천1, 변형NPS +83.3
보고서 품질 (EIEC 5차원 평균)	3.70	0.90	10점 환산 7.40, 구조화 4.08 최고·산업적합성 3.17 최저
유료 전환 가치	—	—	'유용'이상 83.3%(매우유용, ·유용5, 보통2)

주: 응답은 수요기업 12인 전수 응답(결측 없음)을 기준으로 산출하였다. 평균·표준편차는 5점 척도 응답의 기술통계(표본 표준편차, n−1)이며, 표준편차는 원자료에서 산출한 파생값이다. 변형 NPS는 5점 척도 응답을 기반으로 산출한 지표로, 표준 NPS(0−10 척도)와 직접 비교할 수 없다. 보고서 품질은 수요기업이 평가한 EIEC 5차원(구체성 3.92, 실행가능성 3.33, 산업적합성 3.17, 구조화 4.08, 추적가능성 4.00)의 응답자 평균이다. 본 표본은 6장 2절의 전문가 검증(n=15)과 독립된 별도 표본으로 두 결과는 통합하지 않는다. 응답자 3인은 기업명을 기재하지 않았으나 전 문항에 응답을 완료하여 분석에 포함하였다.



[그림 7-14] 수요기업 주요 평정 지표(만족·도입·추천) 평균 (n=12)

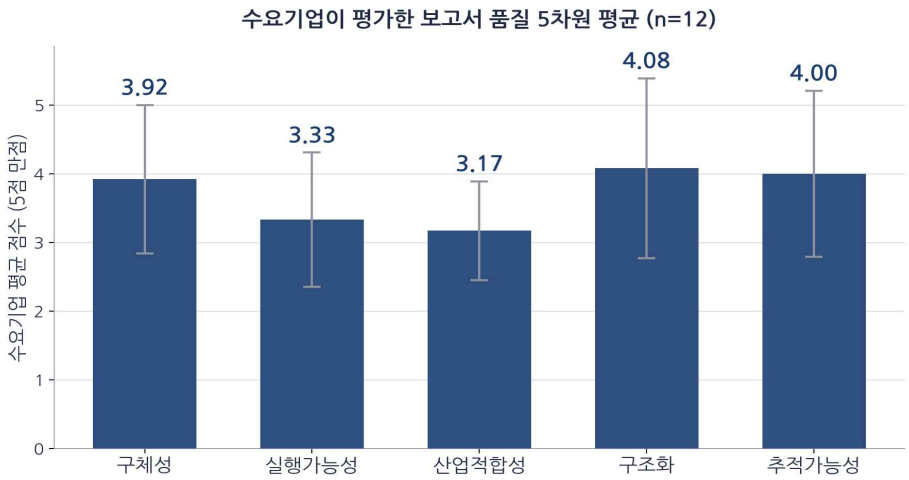
본 절의 결과 수치는 수요기업 ESG 실무 담당자를 대상으로 한 익명 온라인 설문을 통해 확보되었다(n=12, 전수 응답, 결측 없음). 응답 기업의 업종은 제조업 7곳, 건설업 2곳, 서비스업·정보통신(ICT)·기타 각 1곳으로 제조업 중심이었으며, 기업 규모는 종업원 50인 미만이 다수로 10−49인 구간이 6곳으로 가장 많았다. 응답자의 ESG 업무 경력은 1−3년이 7명으로 최다였

고, 대표·부서장·팀장 등 실무 의사결정자와 ESG 전담 인력이 고르게 포함되어 산출물의 실무 활용 적합성을 판단하기에 적절한 표본 구성을 갖추었다. 본 표본은 6장 2절의 전문가 검증(n=15)과 평가 주체가 상이한 독립된 별도 표본이며, 두 결과는 단일 지표로 통합하지 않는다. 또한 본 설문은 산출물의 실무 수용성과 만족도를 측정하기 위한 것으로, 4장, 7장에서 보고한 Agentic 파이프라인의 정량 성능 지표와는 측정 목적과 대상이 다르다.

전반적 만족도는 5점 척도 평균 4.25점으로, 만족(4점 이상) 응답이 83.3%(10/12), 그중 매우 만족(5점)이 50.0%를 차지하였다. 산출물의 실제 도입(활용) 의향은 평균 4.08점으로 적극 도입 의향(4점 이상)이 66.7%, 긍정적 의향(3점 이상)이 100%로 나타났다. 추천 의향은 평균 4.17점이었으며, 추천 응답이 비추천을 크게 상회하여(12명 중 추천 11, 비추천 1, 변형 NPS +83.3), 수요기업 응답자 대다수가 산출물을 추천할 의향을 보였다. 세 평정 지표의 분포는 [그림 7-14]에 제시하였다.

수요기업이 평가한 보고서 품질은 EIEC 5차원 평균 3.70점(5점 만점)으로 10점 환산 시 7.40점이었다. 차원별로는 구조화 4.08점과 추적가능성 4.00점이 상대적으로 높았고, 산업적합성 3.17점과 실행가능성 3.33점이 낮아, 산출물의 체계적 구성과 근거 제시는 긍정적으로 평가된 반면 업종 맞춤화와 현장 실행 가능성에서 개선 여지가 확인되었다([그림 7-15]). 이 7.40점은 수요기업이 실무 활용 관점에서 평가한 값으로, 6장 2절에서 ESG 전문가 15인이 품질 관점에서 평가한 EIEC 점수(6.96/10)와는 평가 주체 및 기준이 다른 독립 지표이므로 상호 합산하거나 직접 비교하지 않는다.

산출물의 유료 전환 가치에 대해서는 '유용' 이상으로 평가한 응답이 83.3%(매우 유용 5명·유용 5명·보통 2명)였고, 현재 산출물을 자사 ESG 대응에 제공할 수 있는 수준인지를 묻는 문항에서는 소폭 수정 이내로 활용 가능하다는 응답이 91.7%(현행 사용 2명·소폭 수정 9명·대폭 수정 1명)에 달하여, 산출물이 추가 가공 부담이 작은 실무 활용 가능 상태로 제공됨을 확인하였다. 한편 유료 전환 가치와 산출물 제공 가능성은 평정척도가 아닌 비율형 문항이므로 그림 없이 [표 7-10]으로만 제시하였다.



[그림 7-15] 수요기업이 평가한 보고서 품질 5차원 평균 (n=12)

자유 응답으로 수집된 정성 피드백은 긍정 의견과 개선 의견의 두 축으로 종합된다. 긍정 측면에서는 ESG 대응의 출발점을 잡기 어려운 중소기업이 우선 개선 과제와 증빙 서식을 일목요연하게 확보할 수 있어 막연한 부담을 줄여 준다는 평가가 다수였다. 개선 측면에서는 업종·평가기관별 기준의 구체화와 동종업계 사례 보강에 대한 요청과 함께, 산출물에 포함된 정량 수치(에너지 사용량·온실가스 배출 등)가 현업에 그대로 적용되기 위해서는 단위·산정 기준의 현업 정합성에 대한 추가 검증이 필요하다는 의견이 제시되었다. 이러한 개선 의견은 향후 업종 특화 지식의 확장과 정량 산출 모듈의 검증 강화 과제로 반영한다.

제 7 절 실증 결과 종합 정량 분석

[표 7-11]은 본 시스템의 성과가 단일 LLM 호출이 아니라 Agentic 다단계 파이프라인 구조에서 비롯되었음을 기존 실험 결과로 보여준다. 첫째, 맥락 주입 단계를 독립된 Agent로 분리하였을 때 동일한 Sonnet 모델 조건에서 컨설팅 품질 총점은 Baseline 7.69에서 RAG+KG 8.52로 0.83점 상승하였으며, 효과크기 d=1.04는 이 차이가 큰 수준임을 뒷받침한다. 둘째, 지식그래

프(KG) 채널을 독립 단계로 배치한 결과 추적가능성은 4.13에서 9.57로 순증 +5.44에 이르러, 단계 분리가 산출물의 근거 추적 역량을 결정적으로 강화함을 확인할 수 있다. 셋째, RAG+KG 조건에서 경량 모델인 Haiku의 점수 8.14가 Baseline 조건에서 고성능 모델인 Sonnet의 점수 7.69를 상회한다는 점은, 경량 모델에 충분한 맥락을 공급하는 구조가 고성능 모델에 빈약한 맥락을 공급하는 구성보다 우수함을 의미하며, 이는 구조적 기여가 모델 성능 격차를 상회할 수 있음을 시사한다. 한편 모델 효과를 검정한 $t=-2.95$($p=0.004$)와 업종별 차이를 검정한 $F=7.26$($p<0.001$)은 구조 효과와 모델 효과를 분리하여 측정한 통계적 근거를 제공한다. 요컨대 본 시스템의 성능 향상은 모델 자체의 역량보다 맥락 주입과 지식그래프 연계를 단계적으로 구성한 Agentic 파이프라인 구조에 상당 부분 기인한다.

[표 7-11] 단일 LLM 호출 대비 Agentic 파이프라인 구조 효과

비교 축	Baseline(단일)	Agentic(RAG+KG)	효과
컨설팅 품질 총점 (Sonnet)	7.69	8.52	+0.83 (d=1.04)
추적가능성	4.13	9.57	순증+5.44
구조 우위 (경량 모델 대비)	7.69 (Baseline 조건, Sonnet)	8.14 (RAG+KG 조건, Haiku)	구조 효과가 모델 격차를 상회

본 장의 실증 결과를 종합하면 다음과 같다. 첫째, 맥락 강화의 효과이다. 3개 맥락 조건 간 품질 총점의 차이는 통계적으로 유의하였으며($F=7.43$, $p<0.01$, $\eta^2=0.146$), Sonnet 기준 Baseline 7.69에서 RAG 8.25(+0.56), RAG-KG 결합 8.52(+0.83)로 단계적으로 상승하였다. RAG에서 RAG-KG 결합으로의 총점 추가 상승(+0.27점)은 유의하지 않았으나(Tukey $p=0.44$), KG의 기여는 추적가능성 차원에서 두드러져(RAG → RAG-KG 결합 +1.47점) Baseline 4.13점에서 RAG-KG 결합 9.57점으로 순증 5.44점에 달하는 전체 5차원 중 가장 큰 개선이 확인되었다. 둘째, 품질·연산 효율성이다. 실무 배포의 최적 조건으로 RAG-KG 결합 조건의 Haiku 모델(8.14점, 비용 3.2배, 비용효율 최적)와 RAG-KG 결합 조건의 Sonnet 모델(8.52점, 비용 45.0배, 최고 품질)의 이원적 배포 전략이 도출되었다. 셋째, HITL 신뢰도의 이원

화이다. 전문가 평가에서 구조화가 AI 채점과 가장 일치하는 차원으로(b3_ai_match=40.0%, 6/15, 1위), 실행가능성이 HITL 개입이 가장 필요한 차원으로(66.7%, 1위) 지목되어 선별 검수 원칙의 정량적 근거를 확보하였다. 넷째, 업종별 차별성이다. 업종 간 품질 차이는 매우 유의하며(F=7.26, p=0.0003, η^2 =0.60), 산업적합성 차원에서 제조업(6.25)과 도소매·기타(0.00) 간의 극명한 차이가 업종 특화 지식 반영의 직접 증거로 확인되었다. 다섯째, AI 재현성이다. 동일 조건 3회 반복 생성의 품질 총점 CV=1.6%로 높은 안정성을 확인하였다.

HANSUNG UNIVERSITY

제 8 장 결 론

제 1 절 연구요약

본 연구는 Agentic AI 기반 멀티 Agent 파이프라인을 통해 중소기업 ESG 맞춤형 컨설팅 방법을 설계구현 및 검증하였다. 이론적으로는 Agentic AI 5단계 프레임워크(Perceive, Reason, Plan, Act, Learn) 중 Stage 1~3을 핵심 구현 범위로 설정하고, 세 개의 협력 Agent(ESG 등급 예측, LLM 기반 ESG 컨설팅, 전문가 검증)로 구성된 멀티 Agent 파이프라인을 중심으로 세 가지 연구 내용에 대한 실증적 해답을 도출하였다. 실증적으로는 비상장 중소기업 299개사의 자가진단 30문항 응답과 외부에서 독립적으로 산정된 ESG 등급을 데이터로 활용하였으며, 180건의 LLM 컨설팅 산출물과 ESG 전문가 15인의 평가를 통해 다각적으로 검증하였다. 특히 이상의 설계는 자가진단 입력부터 산출물 생성까지를 단일 온라인 서비스(esg-platform-v8)로 구현하여, 제안 방법이 개념에 머무르지 않고 실제 동작하는 시스템으로 작동함을 확인하였다.

세 가지 연구 내용에 대한 핵심 결과는 다음과 같다. 첫째, ESG 등급 예측 Agent는 비상장 중소기업의 자가진단 응답만으로 ESG 등급을 실무적으로 유의미한 수준에서 예측할 수 있으며, 자가진단 평균과 파생 정량신호를 더한 맥락에서 예측 성능이 향상됨을 확인하였다. 둘째, LLM 기반 ESG 컨설팅 Agent는 RAG와 KG를 결합한 듀얼 채널 맥락 강화를 통해 컨설팅 산출물의 품질이 통계적으로 유의하게 향상됨을 3×2 완전교차 실험으로 검증하였고, 품질·연산 효율성 분석을 통해 대량 처리용과 프리미엄용으로 구분되는 목적별 최적 배포 조건을 도출하였다. 셋째, 전문가 15인의 HITL 평가를 통해 산출물의 품질과 차원별 신뢰도, 그리고 인간 검수가 필요한 영역을 다면적으로 확인하였다. 각 결과의 구체적 수치와 통계 검정은 4장부터 7장에 상세히 제시하였다.

시스템 도입의 실무적 효과는 시간·비용 절감에서 분명하게 나타난다. 외부 전문 ESG 컨설팅은 1건당 수백만 원 수준의 비용과 사전교육, 현장실사,

보고서 작성에 이르는 수일~수주의 기간을 요구하는 것으로 보고된다.[35] 반면 본 연구가 구현한 시스템은 약 35분 이내에 진단 결과, 개선 과제, 증빙 서식까지의 초기 컨설팅 산출물을 즉시 제공하므로, 유료 컨설팅 대비 초기 진입 비용과 시간을 크게 낮출 수 있다. 다만 본 시스템의 이용 단가는 상용화 이전 단계로 미정이며, 위 비교는 정식 ESG 평가·컨설팅의 대체가 아니라 그 사전 준비 단계를 보완하는 도구 관점에서 해석되어야 한다.

본 연구의 세 연구 내용은 독립적 결과의 병렬이 아니라 단일 파이프라인을 따라 정량 정보가 전이되는 유기적 연결 구조를 이룬다. 출발점에서 ESG 등급 예측 Agent는 자가진단 30문항만으로 정확도 0.8093을, 파생 정량 신호 7개를 추가한 맥락에서 정확도 0.8631(F1 0.8544)을 달성하였으며, 그 출력은 단일 등급 레이블이 아니라 등급, 확률분포, 취약 영역 SHAP의 다차원 정보로 구성되어 후속 컨설팅의 입력 맥락에 직접 투입된다. 이 다차원 예측 정보는 RAG와 KG를 결합한 듀얼 채널 맥락 강화와 진단 요약, 맞춤형 컨설팅 보고서, 증빙 서식의 산출물 3종을 거쳐 컨설팅 산출로 변환되며, 그 효과는 3×2 완전 교차 실험에서 통계적으로 확인되었다(맥락 강화 $F=7.43$, $p<0.01$, $\eta^2=0.146$; Sonnet 기준 품질 총점 7.69 → 8.25 → 8.52). 특히 KG의 연쇄 추론은 추적가능성 차원을 4.13에서 9.57로 순증 5.44만큼 끌어올려, 예측 단계에서 식별된 취약 영역의 근거가 컨설팅 산출까지 추적 가능한 형태로 전달됨을 정량적으로 뒷받침한다. 나아가 업종 간 품질 차이가 매우 유의하고($F=7.26$, $p=0.0003$, $\eta^2=0.60$) 산업적합성 차원에서 제조업 6.25와 도소매·기타 0.00의 극명한 격차가 확인되어, 예측 단계의 업종 정보가 컨설팅 단계의 업종 특화 지식 반영으로 일관되게 연결됨이 입증된다. 끝으로 동일 조건 3회 반복 생성의 품질 총점 변동 계수가 CV=1.6%로 확인되어, 예측에서 컨설팅을 거쳐 검증으로 이어지는 산출의 재현성이 파이프라인 전 구간에서 안정적으로 유지됨을 보인다.

아울러 산출물의 실무 수용성은 수요기업을 대상으로 한 만족도 검증에서도 높게 확인되었다. 5점 척도로 측정한 결과 전반적 만족도는 평균 4.25점, 시스템 도입 의향은 평균 4.08점, 동종 업계 추천 의향은 평균 4.17점으

35) 중소벤처기업진흥공단, 2021; 중소기업중앙회, 2022

로 나타났다. 추천 의향에서는 추천 응답이 비추천 응답을 크게 상회하였고, 이는 산출물의 실무 수용성이 높음을 뒷받침한다. 따라서 수요기업 만족도 검증 결과 본 시스템의 산출물이 실제 중소기업 실무 현장에서 수용될 만한 수준임을 시사한다.

제 2 절 학술적 및 실무적 기여

1. 학술적 기여

첫째, 본 연구는 공시 사각지대에 놓인 비상장 중소기업의 ESG 등급을 자가진단 응답만으로 예측할 수 있음을 실증하였다. 기존 ESG 등급 예측 연구가 상장 대기업의 정형 공시 데이터에 의존해 온 것과 달리, 본 연구는 비상장 중소기업 299개사의 자가진단 30문항 응답과 외부에서 독립적으로 산정된 ESG 등급을 결합하여, 자가진단 단독 입력만으로도 정확도 0.8093의 예측이 가능함을 보였다. 나아가 자가진단 평균과 파생 정량신호를 더한 맥락에서 정확도 0.8631, F1 0.8544로 성능이 향상되어, 저비용·저부담의 입력 데이터만으로 실무적으로 유의미한 등급 예측이 성립함을 입증하였다. 이는 ESG 평가의 혜택이 상장 대기업에 편중되어 온 데이터 격차를 해소하는 방법론적 기여이다.

둘째, 본 연구는 RAG와 지식그래프(KG)를 결합한 듀얼 채널 맥락 강화가 LLM 컨설팅 품질에 미치는 효과를 3×2 완전교차 실험으로 분리 측정하여 그 인과를 통계적으로 규명하였다. 맥락 강화는 품질 총점을 Baseline 대비 +0.83점(10점 만점) 향상시켰으며(ANOVA $F=7.43$, $p<0.01$, $\eta^2=0.146$, 대효과), RAG 단독(+0.56점, $d=0.62$)과 RAG+KG 결합(+0.83점, $d=1.04$)의 효과를 층화하여 측정함으로써 검색 기반 맥락과 구조적 지식 추론의 기여를 독립적으로 식별하였다. 특히 추적가능성 차원에서 KG의 독립적 기여(RAG → RAG+KG +1.47점, Baseline 대비 순증 5.44점)가 전체 5개 차원 중 가장 큰 개선으로 확인되어, KG의 연쇄 추론이 단순 검색을 넘어 컨설팅 근거의 추적가능성에 기여함을 실증하였다. 이는 LLM 응용 연구에서 흔히 정성적으

로만 주장되던 맥락 강화의 효과를 정량적 인과 증거로 전환한 기여이다.

셋째, 본 연구는 업종 특화 지식의 반영이 컨설팅 품질에 미치는 효과를 정량적으로 입증하였다. 업종 간 품질 차이는 통계적으로 매우 유의하였으며 (ANOVA F=7.26, p=0.0003, η^2 =0.60, 대효과), 산업적합성 차원에서 제조업(6.25)과 도소매·기타(0.00) 간의 극명한 차이가 업종 특화 KG가 컨설팅에 실제로 반영되었음을 보여주는 직접 증거로 확인되었다. 이러한 결과는 본 연구의 기여가 단일 LLM의 능력에 있지 않고, 자가진단 → 예측 → 컨설팅 → 검증으로 순환하는 멀티 Agent 구조에 있음을 뒷받침한다. 개별 기반 모델이 빠르게 대체되는 환경에서도 역할이 분리된 Agent 협업, 업종 지식의 구조적 주입, 그리고 인간 검수의 환류라는 설계 원리는 재사용 가능한 학술적 자산으로 남는다.

넷째, 본 연구는 LLM 산출물의 신뢰도를 차원별로 분해하여 검증하는 HITL 평가 프로토콜을 제시하였다. ESG 전문가 15인의 평가에서 AI 자동 채점과 가장 일치하는 차원은 구조화(40.0%, 6/15, 1위)였고, 인간 검수(HITL) 개입이 가장 필요한 차원은 실행가능성(66.7%, 1위)으로 압도적으로 지목되어, 차원별 신뢰도의 이원화가 정량적으로 드러났다. 이는 모든 산출물을 일괄 검수하는 대신 신뢰도가 낮은 차원만 선별적으로 검수하는 선별 검수(selective review) 원칙의 근거를 제공하며, LLM-as-judge 자동 평가와 인간 전문가 평가를 결합한 검증 설계의 학술적 사례가 된다. 또한 동일 조건 3회 반복 생성의 품질 총점 변동계수가 CV=1.6%로 확인되어, 본 파이프라인의 산출 안정성이 정량적으로 뒷받침된다.

2. 실무적 기여

첫째, 본 연구는 전문 ESG 컨설팅의 지리적·시기적 한계를 해소하는 상시 접근형 서비스를 실제로 구현하였다. 전문 컨설팅이 수도권에 집중되고 연간 용량이 한정된 반면, 본 시스템은 자가진단 입력부터 산출물 생성까지를 단일 온라인 서비스(esg-platform-v8)로 24시간 제공하여, 지방·소규모 중소기업도 동등한 진단과 개선 방향을 확보할 수 있다. 이는 전문 컨설팅의 완전

한 대체가 아니라 초기 진단 단계의 구조적 보완재로 기능하며, 고비용 전문 컨설팅이 필요한 기업과 진입 비용이 장벽인 다수 중소기업으로 시장을 이원화하는 합리적 배포 전략의 근거가 된다.

둘째, 본 연구는 품질과 연산 비용을 함께 고려한 목적별 최적 배포 조건을 정량적으로 제시하였다. 분석 결과, RAG+KG 조건의 Haiku 모델 조합(8.14점, 비용 3.2배)은 비용 효율이 가장 우수하여 대량 배포에 적합하고, RAG+KG 조건의 Sonnet 모델 조합(8.52점, 비용 45배)은 최고 품질을 제공하여 프리미엄 컨설팅에 적합한 이원적 전략이 도출되었다. 이는 서비스 운영자가 처리 규모와 품질 요구 수준에 따라 모델을 선택할 수 있는 구체적 의사결정 지침을 제공한다.

셋째, 본 시스템은 진단 결과를 실무자가 곧바로 실행할 수 있는 형태의 산출물로 전환한다. 자가진단 응답은 예측 등급과 취약 영역으로 변환되고, 이는 다시 맞춤형 컨설팅 보고서로 이어지며, 컨설팅 결과가 워드 증빙 서식 5종으로 자동 생성되어 ESG 담당자가 별도 문서 작성 부담 없이 즉시 실행에 옮길 수 있도록 설계되었다. 이는 진단에 그치지 않고 실행으로 연결되는 실무적 완결성을 확보한 기여이다. 본 연구의 시스템 구성은 박사학위 연구의 산출물로서 특허 출원(출원번호 10-2026-0089309, 2026년 5월 18일 출원, 발명자 노은미·노광현, 청구항 15개)으로 이어졌다. 자가진단 응답으로부터 파생 정량신호를 산출하여 ESG 등급을 예측하고 이를 검색증강생성 및 지식그래프 기반 컨설팅 생성과 연계하는 구조의 신규성과 산업적 실용성을 외부 절차를 통해 확인한 결과이다.

제 3 절 향후 연구 방향

본 연구의 성과는 동시에 후속 연구로 확장될 수 있는 지점을 함께 제시한다.

첫째, 표본의 대표성이다. 본 연구의 분석 대상은 ESG 자가진단에 자발적으로 참여한 299개사로 구성되어 ESG에 관심이 높은 기업이 상대적으로 많이 포함되었을 수 있으며, 응답 기업이 1~2개사에 불과한 업종(서비스업·도

소매업 등)도 있었다. 따라서 전국 단위의 무작위·층화 표집과 해당 업종 및 High 등급 기업의 추가 확보를 통해 일반화 가능성을 한층 높일 수 있다. High 등급은 전체의 5.0%(15개사)로 상대적으로 희소하여, 본 연구는 교차검증의 각 분할에서 등급 비율을 동일하게 유지하는 Stratified K-Fold 기법으로 이를 보정하였으며, 후속 연구에서 해당 등급 표본을 보강하면 예측 정확도를 더욱 안정화할 수 있다.

둘째, 종단적 인과 효과의 규명이다. 본 연구가 횡단적 설계로 시스템의 기술적 타당성과 실무 수용성을 확인한 데 이어, 컨설팅을 수신한 기업을 일정 기간 추적하는 패널 조사를 통해 ESG 등급 변화의 인과 효과를 검증하는 연구로 발전시킬 수 있다.

셋째, 가장 핵심적으로 프레임워크의 완성이다. 본 연구는 Agentic AI 5단계 프레임워크 가운데 Stage 1~3을 구현·검증하고, 전문가 검증 단계인 Stage 4의 개념적 타당성을 15인의 전문가 평가로 확인하였으며, 모니터링·학습 단계인 Stage 5를 설계 수준에서 제안하였다. 후속 연구의 핵심 과제는 Stage 4에서 Stage 5로의 전이를 통한 오케스트레이션의 완성이다. 구체적으로 전문가 검수 결과를 상위 오케스트레이터가 수집·정량화하여 ESG 등급 예측 Agent와 LLM 기반 ESG 컨설팅 Agent의 개선 근거로 자동 환류시키는 폐루프(closed-loop) 구조를 구현해야 한다. 즉 전문가 피드백 신호를 여섯 구현 모듈(A1~A6)별 개선 우선순위로 변환하고 이를 다음 컨설팅 사이클의 검색, 예측, 생성 과정에 반영하는 학습 루프가 Stage 5의 실체이며, 이때 오케스트레이터는 배포 이후 기업의 ESG 등급 변화를 지속적으로 관측하여 모델 재학습의 입력으로 삼는다. 이를 통해 본 연구가 기본 구현(baseline implementation) 수준에서 제시한 Agentic AI 파이프라인은 전문가 검증과 종단적 관측이 예측·컨설팅 단계로 환류되는 자율적 개선 구조로 발전할 수 있다.

참 고 문 헌

1. 국내문헌

관계부처 합동 (2021). K-ESG 가이드라인 v1.0. 산업통상자원부.

관계부처 합동 (2022). K-ESG 가이드라인 v2.0. 산업통상자원부.

고용노동부 (2025). 2024년 산업재해 현황분석. 고용노동부 산재예방정책국.

공정거래위원회 (2024). 공정거래 자율준수프로그램(CP) 운영 가이드라인. 공정거래위원회.

금융위원회 (2023). ESG 금융 추진단 제3차 회의 자료 — ESG 공시 로드맵. 금융위원회.

김현수, 이지은, 박정호 (2022). 중소기업 ESG 진단 방법론 개발: 자가진단 체계 중심으로. 중소기업연구, 44(2), 137-158.

박준영·박준형·허원진·양성병 (2025). LLM을 활용한 ESG 평가 기준 데이터 추출 및 데이터베이스 구축. 인터넷전자상거래연구, 25(1), 163-180. https://doi.org/10.37272/JIECR.2025.2.25.1.163

산업통상자원부 (2024). K-ESG 가이드라인 개정판. 산업통상자원부. (2024년 12월 발간)

산업통상자원부 (2024). 업종별 K-ESG 가이드라인(반도체·석유화학·철강·자동차·디스플레이). 산업통상자원부. (2024년 12월 발간)

산업통상자원부 (2024). K-ESG 자가진단 시스템 운영 보고서. 산업통상자원부.

산업통상자원부·한국생산성본부 (2024). 공급망 대응 K-ESG 가이드라인 v2.0. 산업통상자원부. (2024년 12월 발간)

이연우 (2022). 한국형 ESG 가이드라인의 쟁점분석과 적용포인트.

Auditor Journal, 266호.

정재학 (2024). ESG 기업 간 보고서 사례를 중심으로 LLM RAG를 활용한 평가에 관한 연구. 서강대학교 대학원 경영학과 석사학위논문.

제미화 (2024). 지속가능경영보고서 발간 기업 간 네트워크 분석: Large Language Model을 활용한 의미적 유사도를 중심으로. 서강대학교 대학원 경제학과 석사학위논문.

중소기업중앙회 (2023). 중소기업 ESG 대응 현황 실태조사 보고서. 중소기업중앙회 경제정책본부, 연구보고서 2023-07.

중소기업중앙회 (2023). 중소기업 ESG 경영 실천을 위한 ESG 규정례 자료집. 중소기업중앙회.

중소기업진흥공단 (2022). 중소기업 ESG 자가진단 도구 개발 및 보급 사업 보고서. 중소기업진흥공단 정책연구실.

중소벤처기업부 (2022). 중소기업 ESG 경영 인식 및 대응 실태 조사 보고서. 중소벤처기업부 정책연구보고서 2022-04.

중소벤처기업부·중소벤처기업진흥공단 (2023). 중소벤처기업을 위한 ESG 경영안내서 — 공급망편. 중소벤처기업부 산업정책관실.

중소벤처기업부·중소벤처기업진흥공단 (2023). 중소벤처기업을 위한 ESG 경영안내서 — 이해편. 중소벤처기업부 산업정책관실.

최은실 (2025). 생성형 AI를 활용한 ESG 서류검토 자동화 시스템의 개발. 고려대학교 SW·AI융합대학원 빅데이터융합학과 석사학위논문.

한국경영학회 (2024). ESG 경영과 중소기업: 정책·전략·실무. 한국경영학회 ESG 연구회 편, 박영사.

한국ESG기준원 (2023). ESG 평가 방법론 및 기준 — 2023년 개정판. 한국ESG기준원.

2. 국외문헌

Acharya, D. B., Kuppan, K., & Divya, B. (2025). Agentic AI: Autonomous intelligence for complex goals – A comprehensive survey. IEEE Access, 13, 18912–18936.

Anthropic (2024a). Building effective agents. Anthropic Engineering. https://www.anthropic.com/engineering/building-effective-agents

Anthropic (2024b). Claude 3 model card and technical documentation. Anthropic Technical Report.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Zhang, J., & Zhao, Z. (2025). Corporate ESG rating prediction based on XGBoost-SHAP interpretable machine learning model. Expert Systems with Applications, 295, Article 128809. https://doi.org/10.1016/j.eswa.2025.128809

Zou, Y., Shi, M., Chen, Z., Deng, Z., Lei, Z., Zeng, Z., Yang, S., Tong, H., Xiao, L., & Zhou, W. (2023). ESGReveal: An LLM-based approach for extracting structured data from ESG reports. arXiv preprint arXiv:2312.17264. https://arxiv.org/abs/2312.17264

Cai, B., Ye, Z., & Chen, S. (2025). Intelligent ESG evaluation for construction enterprises in China: An LLM-based model. Buildings, 15(15), Article 2710. https://doi.org/10.3390/buildings15152710

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).

ACM.

Chowdhury, S., Rahman, M., & Kim, J. (2023). Global ESG rating prediction for listed companies across 73 countries: A machine learning comparison. Journal of Sustainable Finance & Investment, 13(4), 412–438.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Lawrence Erlbaum Associates.

Eccles, R. G., Ioannou, I., & Serafeim, G. (2014). The impact of corporate sustainability on organizational processes and performance. Management Science, 60(11), 2835–2857.

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., & Larson, J. (2024). From local to global: A graph RAG approach to query-focused summarization. arXiv preprint arXiv:2404.16130.

Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). RAGAS: Automated evaluation of retrieval augmented generation. In Proceedings of EACL 2024 (pp. 150–158). Association for Computational Linguistics.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. Journal of Marketing Research, 18(1), 39–50.

Friede, G., Busch, T., & Bassen, A. (2015). ESG and financial performance: Aggregated evidence from more than 2,000 empirical studies. Journal of Sustainable Finance & Investment, 5(4), 210–233.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.

Gartner (2025). Predicts 2026: Agentic AI will reshape enterprise software. Gartner Research.

He, C., Zhou, X., Wu, Y., et al. (2025). ESGenius: Benchmarking LLMs on Environmental, Social, and Governance (ESG) and Sustainability Knowledge. In Proceedings of EMNLP 2025 (pp. 14612–14653). arXiv:2506.01646

Hogan, A., Blomqvist, E., Cochez, M., D'Amato, C., Melo, G. de, Gutierrez, C., et al. (2021). Knowledge graphs. ACM Computing Surveys, 54(4), 1–37.

Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Zhang, C., et al. (2024). MetaGPT: Meta programming for a multi-agent collaborative framework. In Proceedings of ICLR 2024 (pp. 1–20).

IBM Think (2025). What is human-in-the-loop? IBM Think. https://www.ibm.com/think/topics/human-in-the-loop

Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. Journal of Consumer Research, 30(2), 199–218.

Jiang, X. (2024). Predicting corporate ESG scores using machine learning: A comparative study. Proceedings of the 3rd International Conference on Financial Technology and Business Analysis. https://doi.org/10.54254/2754-1169/118/2024.18728

Karakurt, E., & Akbulut, A. (2026). Retrieval-augmented generation (RAG) and large language models (LLMs) for enterprise knowledge management and document automation: A systematic literature review. Applied Sciences, 16(1), 368. https://doi.org/10.3390/app16010368

Khamis, A., Elshakankiri, M., & Elsayed, H. (2025). Agentic AI Systems: Architecture and Evaluation Using a Frictionless Parking Scenario. IEEE Access, 13, 126052-126069. https://doi.org/10.1109/ACCESS.2025.3590264

Kirkpatrick, D. L. (1959). Techniques for evaluating training programs. Journal of ASTD, 13(11), 3-9.

Kirkpatrick, D. L. (1994). Evaluating Training Programs: The Four Levels. San Francisco: Berrett-Koehler Publishers.

LangChain (2024). State of AI agents report 2024. LangChain Blog.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

Liu, W. (2025). Research on the construction and application of retrieval enhanced generation (RAG) model based on knowledge graph. Scientific Reports, 15, Article 40425. https://doi.org/10.1038/s41598-025-21222-z

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765-4774.

OpenAI (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774.

OpenAI (2024). GPT-4o system card. OpenAI Technical Report.

Parseur (2025). Human-in-the-loop AI (HITL): Complete guide to benefits, best practices and trends for 2026. Parseur Blog. https://parseur.com/blog/human-in-the-loop-ai

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Porter, M. E. (1996). What is strategy? Harvard Business Review, 74(6), 61-78.

Sinek, S. (2009). Start with Why: How Great Leaders Inspire Everyone to Take Action. New York: Portfolio/Penguin.

Singla, A., Sukharevsky, A., Yee, L., Chui, M., & Hall, B. (2025). The state of AI: How organizations are rewiring to capture value. McKinsey & Company. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-how-organizations-are-rewiring-to-capture-value

SuperAnnotate (2025). LLM-as-a-judge vs. human evaluation: Why together is better. SuperAnnotate Blog. https://www.superannotate.com/blog/llm-as-a-judge-vs-human-evaluation

Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley.

UN Global Compact (UNGC) (2004). Who cares wins: Connecting financial markets to a changing world. United Nations Global Compact Report.

UN Global Compact (UNGC) (2023). SME ESG framework and self-assessment toolkit. UN Global Compact SME Engagement Initiative.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., et al. (2024). A survey on large language model based autonomous agents.

Frontiers of Computer Science, 18(6), Article 186345.

World Economic Forum (WEF) (2020). Measuring stakeholder capitalism: Towards common metrics and consistent reporting of sustainable value creation. WEF White Paper.

World Economic Forum (WEF) (2023). ESG metrics for sustainable value creation. World Economic Forum Report.

Wu, Q., Xiang, X., Huang, H., et al. (2025). SusGen-GPT: A Data-Centric LLM for Financial NLP and Sustainability Report Generation. In Findings of NAACL 2025. arXiv:2412.10906

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., et al. (2023). The rise and potential of large language model based agents: A survey. arXiv preprint arXiv:2309.07864.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate Data Analysis (8th ed.). Cengage Learning.

Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. Journal of the Academy of Marketing Science, 43(1), 115-135.

Karpathy, A. (2025). There's a new kind of coding I call vibe coding [Post]. X (formerly Twitter).

Principles for Responsible Investment (PRI) (2006). Principles for Responsible Investment. UNEP Finance Initiative & UN Global Compact.

European Union (2014). Directive 2014/95/EU (Non-Financial Reporting Directive). Official Journal of the European Union.

European Union (2022). Directive (EU) 2022/2464 (Corporate Sustainability Reporting Directive). Official Journal of the European Union (in force 2024).

3. 데이터 및 온라인 자료

Anthropic (2024). Claude API documentation. https://docs.anthropic.com

Google (2026). Google Colaboratory. https://colab.research.google.com

GitHub (2026). GitHub Repository Hosting. https://github.com

Vercel (2026). Vercel Deployment Platform. https://vercel.com

개인정보보호위원회 (2024). 개인정보보호 교육자료 및 온라인 교육. https://www.pipc.go.kr

고용노동부 (2024). 고용노동부 홈페이지. https://www.moel.go.kr

공정거래위원회 (2024). 공정거래 자율준수프로그램(CP) 안내. https://www.ftc.go.kr

금융감독원 (2024). 전자공시시스템(DART). https://dart.fss.or.kr

안전보건공단 (2024). 산업재해 통계 및 안전보건 자료. https://kosha.or.kr

한국에너지공단 (2024). 온실가스 목표관리제 종합정보시스템(TIPS). https://tips.energy.or.kr

온실가스종합정보센터 (2024). 온실가스 배출량 통계. https://gir.go.kr

한국인정지원센터 (2024). 인증기관 및 인증통계. https://kab.or.kr

환경부 (2024). 국가환경정보센터(NCIS). https://ncis.me.go.kr

환경부 (2024). 화학물질정보시스템(ICIS). https://icis.me.go.kr

별 첨 문 서

[별첨 1] 전문가 평가(HITL 검증) 설문내용

연구 개요: 본 설문은 LLM 기반 ESG 컨설팅 시스템이 생성한 보고서의 품질을 ESG 전문가가 평가하여 측정도구(EIEC 5차원)의 신뢰도·타당도를 검증하기 위한 것이다. (한성대학교 대학원 스마트융합컨설팅학과, 연구자 노은미)
평가 대상: ① 자가진단 결과보고서 ② 맞춤형 컨설팅 보고서 ③ 증빙서식 5종
소요 시간: 약 5분 / 익명 처리: 응답은 익명으로 처리되며 박사학위 논문 연구 목적으로만 사용된다.

Part 1. 평가자 정보
1. 성함 (선택): _____
2. 전문 분야 (필수):
① ESG 컨설팅 ② ESG 인증·심사 ③ 정책·연구 ④ AI·데이터 ⑤ 기타
3. ESG 관련 경력 (필수): ① 3년 미만 ② 3~5년 ③ 5~10년 ④ 10년 이상
4. ESG 컨설팅 수행 건수 (필수): ① 10건 미만 ② 10~30건 ③ 30~50건 ④ 50건 이상
5. 주요 컨설팅 대상 (필수):
① 대기업 ② 중견기업 ③ 중소기업 ④ 소상공인·스타트업 ⑤ 다양

Part 2. 보고서 품질 평가 — EIEC 5차원 (각 문항 1~5점)
척도: 1 매우 미흡 / 2 미흡 / 3 보통 / 4 우수 / 5 매우 우수
1. 구체성: 개선 과제가 정량 수치·기한·담당을 구체적으로 제시하는 정도 …… (1 · 2 · 3 · 4 · 5)
2. 실행가능성: 개선 과제가 중소기업의 자원(예산·인력·시간) 범위 내에서 실행 가능한 정도 …… (1 · 2 · 3 · 4 · 5)
3. 산업적합성: 개선 과제가 업종 특성과 업종별 핵심 ESG 이슈를 반영하는 정도 …… (1 · 2 · 3 · 4 · 5)
4. 구조화: 개선 과제가 우선순위·논리 체계를 갖추어 일관되게 제시되는 정도 …… (1 · 2 · 3 · 4 · 5)
5. 추적가능성: 근거(법규·인증기준·모범사례)가 정확히 인용되고 출처를 추적·검증할 수 있는 정도 …… (1 · 2 · 3 · 4 · 5)

Part 3. AI 자동채점 상호검증

B1. 보고서에서 AI가 법률명·조항을 틀리게 인용한 사례가 있습니까? (필수):

① 없음 ② 1건 ③ 2~3건 ④ 4건 이상 ⑤ 판단 불가

B2. 5개 차원 중 AI 자동채점만으로 신뢰하기 어렵고 전문가 검수(HITL)가 반드시 필요한 차원은? (필수):

① 구체성 ② 실행가능성 ③ 산업적합성 ④ 구조화 ⑤ 추적가능성

B3. 반대로 AI 자동채점이 전문가 평가와 가장 일치할 것으로 예상되는 차원은? (필수): ① 구체성 ② 실행가능성 ③ 산업적합성 ④ 구조화 ⑤ 추적가능성

Part 4. 종합 판단

C1. AI 생성 보고서의 전반적 품질은 전문 컨설턴트 대비 어느 수준입니까? (필수): ① 전문가 수준 이상 ② 전문가와 유사 ③ 전문가의 70~80% ④ 전문가의 50~70% ⑤ 50% 미만

C2. 이 보고서를 중소기업에 그대로 제공하기에 적절합니까? (필수):

① 그대로 제공 가능 ② 일부 수정 후 가능 ③ 상당한 수정 필요 ④ 제공 불가

C3. AI + 전문가 HITL 이중검증 구조가 기존 전문 ESG 컨설팅을 어느 정도 대체할 수 있다고 보십니까? (필수): ① 완전 대체 가능 ② 대부분 대체 가능 ③ 부분 대체 가능 ④ 보조 도구로만 활용 ⑤ 대체 불가

C4. AI 기반 ESG 컨설팅 도입 시 가장 큰 장점은? (필수):

① 비용 절감 ② 시간 단축 ③ 일관된 품질 ④ 24시간 접근성 ⑤ 대량 처리

C5. AI 기반 ESG 컨설팅 도입 시 가장 큰 우려는? (필수): ① 법규 정확성 ② 업종 특화 부족 ③ 책임 소재 불명확 ④ 기밀 유출 ⑤ AI 신뢰 부족

C6. 본 연구의 포지셔닝('중소기업이 정식 ESG 평가 전 활용하는 자가진단 보완 도구')에 대한 평가는? (필수):

① 매우 타당 ② 타당 ③ 보통 ④ 부적절 ⑤ 매우 부적절

Part 5. 자유 의견 (선택)

본 연구에 도움될 의견(개선 제안·우려 사항·연구 방향 등)을 자유롭게 적어주십시오.

_____

[별첨 2] ESG 수요기업 검증 설문내용

연구 개요: 본 설문은 LLM 기반 ESG 컨설팅 시스템의 산출물에 대한 수요기업(중소기업)의 만족도와 실무 수용성을 검증하기 위한 것이다.
(한성대학교 대학원 스마트융합컨설팅학과, 연구자 노은미)
평가 대상: ① 자가진단 결과보고서 ② 맞춤형 컨설팅 보고서 ③ 증빙서식 5종
소요 시간: 약 5분 / 익명 처리: 응답은 익명으로 처리되며 박사학위 논문 연구 목적으로만 사용된다.

Part 1. 기업 정보
1. 회사명 (선택): _____
2. 업종 (필수):
① 제조업 ② 정보통신(ICT) ③ 도소매·유통 ④ 건설업 ⑤ 서비스업 ⑥ 기타
3. 회사 규모(상시 근로자 수) (필수): ① 10인 미만 ② 10~49인 ③ 50~99인 ④ 100~299인 ⑤ 300인 이상
4. 직책·담당 업무 (선택): _____
5. ESG 업무 경험 (필수): ① 없음 ② 1년 미만 ③ 1~3년 ④ 3~5년 ⑤ 5년 이상
6. 우리 회사의 ESG 현황 (복수 선택): □ ESG 전담 조직이 있다 □ ESG 평가·실사를 받아본 적이 있다 □ 거래처로부터 ESG 자료 요청을 받은 적이 있다 □ 해당 사항 없음

Part 2. 보고서 품질 평가 (각 문항 1~5점)
척도: 1 전혀 그렇지 않다 / 2 그렇지 않다 / 3 보통 / 4 그렇다 / 5 매우 그렇다
1. 개선 과제가 구체적인가요? 수치·기한·담당이 명확하게 제시되어 있는지 …… (1 · 2 · 3 · 4 · 5)
2. 우리 회사가 실제로 실행할 수 있는 내용인가요? 예산·인력·시간 범위 내에서 실행 가능한지 …… (1 · 2 · 3 · 4 · 5)
3. 우리 업종에 맞는 내용인가요? 업종 특성과 주요 이슈를 반영하는지 …… (1 · 2 · 3 · 4 · 5)
4. 보고서가 체계적으로 정리되어 있나요? 우선순위·분류가 명확하고 이해하기 쉬운지 …… (1 · 2 · 3 · 4 · 5)
5. 근거·출처가 명확한가요? 개선 과제의 근거(법규·기준 등)와 출처가 명확한지 …… (1 · 2 · 3 · 4 · 5)

Part 3. 만족도·활용성

1. 전반적 만족도 (필수, 1~5): 제공된 자료 전반에 대한 만족도 (1 매우 불만족 ~ 5 매우 만족) …… (1 · 2 · 3 · 4 · 5)

2. 도입 의향 (필수, 1~5): 정식 제공 시 우리 회사에서 활용할 의향 (1 전혀 없다 ~ 5 매우 높다) …… (1 · 2 · 3 · 4 · 5)

3. 컨설팅 보고서를 우리 회사 ESG 업무에 활용한다면? (필수): ① 그대로 활용 가능 ② 일부 수정 후 활용 가능 ③ 상당한 수정이 필요 ④ 활용하기 어려움

4. 유료 컨설팅과 비교한 이 자료의 가치는? (필수): ① 매우 유용(유료 못지않다) ② 유용(충분히 참고할 만하다) ③ 보통 ④ 낮음 ⑤ 거의 없음

5. 이 서비스의 가장 큰 장점은? (필수): ① 비용 절감 ② 시간 단축 ③ 무엇부터 해야 할지 방향 제시 ④ 바로 쓸 수 있는 서식 제공 ⑤ 쉬운 접근성

6. 가장 우려되는 점은? (필수): ① 법규 내용의 정확성 ② 업종 특성 반영 부족 ③ 문제 발생 시 책임 소재 ④ 회사 정보 유출 ⑤ AI 신뢰 부족

7. 추천 의향 (필수, 1~5): 다른 회사에 추천하겠습니까? (1 전혀 아니다 ~ 5 매우 그렇다) …… (1 · 2 · 3 · 4 · 5)

Part 4. 자유 의견 (선택)

1. 이 자료가 가장 도움이 될 것 같은 부분은?

_____

2. 부족하거나 아쉬운 부분이 있다면?

_____

3. 추가로 제공되었으면 하는 자료나 기능이 있다면?

_____

HANSUNG UNIVERSITY

ABSTRACT

An LLM-Based Agentic AI Framework for Tailored ESG Consulting in Small and Medium-sized Enterprises

Roh, Eunmi

Major in Smart Convergence Product

Dept. of Smart Convergence Consulting

The Graduate School

Hansung University

ESG(Environmental, Social, and Governance) management has become a baseline requirement across global supply chains. The European Union's Corporate Sustainability Reporting Directive, which took effect in 2024, together with the disclosure standards issued by the International Sustainability Standards Board, has extended reporting obligations to the smaller suppliers embedded within these chains. Korean small and medium sized enterprises in particular face structural barriers of cost and limited in-house expertise, while conventional professional consulting remains costly and time consuming. To lower these barriers without sacrificing expert level quality, this study designs, implements, and validates an agentic artificial intelligence method for customized ESG consulting tailored to smaller firms, in which multiple language model driven agents collaborate within a structured pipeline.

The Agentic AI framework organizes the consulting process into five stages(Perceive, Reason, Plan, Act, and Learn) configured not as a linear sequence but as a closed feedback loop, in which each stage passes structured context to the next and informs earlier stages as needed. A higher-level orchestrator coordinates the flow from self assessment input through grade prediction and customized consulting to expert validation. The first three stages, spanning self assessment to consulting generation, define the implementation and validation scope, while the expert validation stage is presented as a proof of concept.

The Agentic AI framework is realized through three collaborating agents. The first agent predicts an ESG grade, mapping self-assessment responses to one of three classes through a gradient boosting classifier and returning the predicted grade, its class probability distribution, and explanations of vulnerable areas based on Shapley values. The second agent generates consulting content: taking the diagnosis as context, it draws on two complementary channels retrieval augmented generation and a knowledge graph to link relevant regulations, certifications, and best practices, and produces three deliverables, namely a diagnostic summary, a customized consulting report, and an evidence-form guide. The third agent validates these outputs within a human in the loop structure, in which fifteen experts independently rate them across the five quality dimensions of the proposed evaluation framework, producing dimension level scores, measures of inter rater agreement, and structured feedback that returns to the preceding agents. The three agents function not as independent modules but as a single organic pipeline.

This framework was implemented as a fully working, installation free online service that brings together the three agents, the orchestrator, the dual retrieval channels, and a web based self assessment interface, operating end to end from self assessment to output generation. The

system was validated on a sample of 299 unlisted small and medium-sized enterprises through a crossed design of context conditions and models. It predicted ESG grades with an accuracy of 0.8631. This demonstrates that an externally assessed grade can be inferred from self-reported data alone. Combining retrieval-augmented generation with the knowledge graph yielded a peak consulting-quality score of 8.52, confirming the value of context enrichment. Beyond this expert quality assessment, a separate satisfaction survey of twelve small and medium sized enterprises that used the service reported an average overall satisfaction of 4.25 on a five point scale, with 83.3 percent expressing satisfaction, and an average adoption intention of 4.08, confirming the practical acceptability of the deliverables from the standpoint of end users.

The study makes three contributions. It shows empirically that ESG grades for unlisted smaller firms can be predicted from self assessment data alone; it integrates ESG self assessment and automated consulting into a single multi agent pipeline; and it proposes and validates both a five dimension quality evaluation framework and the ESG Bridge Tool concept. By delivering expert level consulting as a low cost online service and using human in the loop evaluation to separate tasks suited to automation from those requiring expert judgment, the study offers a practical basis for real world deployment. Future work includes the full implementation of Stages 4 and 5 and the extension toward a self-improving structure through longitudinal observation.

【Key word】 ESG Consulting; SMEs; Agentic AI; Large Language Model (LLM); Knowledge Graph; Expert Validation; Human in the Loop (HITL)