

Generative AI-Based ITGC Audit Workpaper Automation: Performance Evaluation and Practitioner Acceptance

Jisu Sohn*, Kyuhwan Lee**, YoungGeun Park***, Dongsu Lee****

*Master's Degree, Dept. of Digital Management, Graduate School, Korea University, Sejong, Korea

**Software Developer, Woori Bank, Seoul, Korea

***Ph.D., Dept. of Digital Management, Graduate School, Korea University, Sejong, Korea

****Professor and Director, Corporate Network Center, Gachon University, Incheon, Korea

[Abstract]

Reliability concerns arising from hallucinations and automation bias constrain the adoption of generative AI in auditing. This study proposes an automated model for drafting Information Technology General Controls (ITGC) workpapers using an eight-block prompt and type-safe output. The system was compared with an experienced evaluator using identical evidence across six controls, supplemented by a practitioner survey. In the initial single-run experiment, workpaper drafting time decreased by 96.2%, and agreement with the single reference evaluator reached 98.7% across 154 refined synthetic-evidence cases; because 148 of the 154 cases were PASS, this figure reflects agreement on a highly imbalanced dataset rather than general audit accuracy, and the detection of all four FAIL cases is a preliminary result. No hallucinations were identified under the predefined criterion in the final post-processed outputs. However, agreement decreased to 83.6% in a subsequent rerun of 146 cases across three controls. Three evaluators showed complete agreement on 60 cases (Fleiss' $\kappa = 1.00$). The experienced evaluator performed better in identifying exceptions requiring contextual judgment, while the survey indicated perceived needs for human intervention. These findings are observations under refined synthetic evidence and the specific system configuration examined; they do not establish that individual system components caused the high agreement or the absence of hallucinations, and the effects of human intervention on error reduction and workpaper quality require further validation.

► **Key words:** Generative Artificial Intelligence, Internal Control over Financial Reporting (ICFR), IT General Controls (ITGC), Professional Skepticism, Hallucination, Human-in-the-Loop (HITL)

-
- First Author: Jisu Sohn, Corresponding Author: Dongsu Lee
 - *Jisu Sohn (sohn7654@gmail.com), Dept. of Digital Management, Graduate School, Korea University
 - **Kyuhwan Lee (gw1221@naver.com), Woori Bank
 - ***YoungGeun Park (FL9871@naver.com), Dept. of Digital Management, Graduate School, Korea University
 - ****Dongsu Lee (leedongsu.kr@gmail.com), Corporate Network Center Director, Gachon University
 - Received: 2026. 07. 24, Revised: 2026. 10. 01, Accepted: 2026. 10. 05.

[요 약]

환각과 자동화 편향에 따른 신뢰성 우려는 감사 분야의 생성형 AI 도입을 제약한다. 본 연구는 8-Block 프롬프트와 자료형 안전 출력을 적용한 IT 일반통제(ITGC) 조서 자동화 모델을 제안하고, 6개 통제의 동일 증빙에 대해 숙련 평가자와 비교하며 실무자 설문을 병행하였다. 최초 단회 실험에서 조서 초안 작성 시간은 96.2% 단축되었고, 정제된 가상 증빙 154건에서 단일 기준 평가자와의 판정 일치율은 98.7%였다. 이는 154건 중 148건이 PASS인 불균형 데이터에서의 일치율로 일반적인 감사 정확도를 뜻하지 않으며, FAIL 4건의 전건 탐지는 초기 검증 결과이다. 후처리된 최종 산출물에서 정의된 환각은 식별되지 않았으나, 3개 통제 146건의 재실행 일치율은 83.6%로 낮아졌다. 3인 평가자의 60건 판정은 완전히 일치하였다(Fleiss' $\kappa=1.00$). 숙련 평가자는 맥락이 필요한 예외 식별에 강했으며, 설문은 인간 개입의 필요성에 대한 인식을 보여주었다. 결과는 정제된 가상 증빙과 현재의 시스템 구성이 결합된 조건에서의 관찰로 한정되며, 개별 구성요소가 높은 일치도나 환각 미관찰에 직접 기여하였음을 의미하지 않는다. 인간 개입의 실제 오류 감소 및 품질 개선 효과는 후속 검증이 필요하다.

▶ **주제어:** 생성형 인공지능, 내부회계관리제도, IT 일반통제, 전문적 의구심, 환각 현상, 인간 개입

I. Introduction

생성형 AI는 비정형 데이터의 심층 분석·요약·문서화 등 고차원적 인지 과업으로 수행 범위를 확장하며 전문직의 업무 방식을 변화시키고 있다. 전 세계 AI 지출은 2026년 약 2조 5,900억 달러로 전년 대비 47% 증가할 것으로 전망되며[1], 758명의 전문 컨설턴트를 대상으로 한 실험에서는 생성형 AI가 구조화된 지식 업무의 생산성과 품질을 유의하게 향상시키는 것으로 나타났다[2]. 회계·감사 분야에서도 반복적·규칙 기반 과업의 자동화 가능성이 제시된 이후 실증적 검증으로 확장되어 왔으며[3][4][5][6], 국내에서도 대형 회계법인이 내부회계관리제도(K-SOX) 평가 자동화 솔루션을 실무에 적용하는 등[7] ITGC 평가 자동화는 이미 도입되고 있는 영역이다. 특히 ITGC의 운영 효과성 평가는 다수의 증빙을 반복적으로 대조하고 조서로 문서화하는 노동집약적 과업으로, 평가자의 숙련도에 따라 판단의 일관성이 달라질 수 있어 자동화의 실익이 큰 영역으로 평가된다. 그럼에도 불구하고 감사와 같은 고신뢰 영역에서는 자동화 산출물의 정확성과 신뢰성이 충분히 검증되지 않으면 도입이 제약되므로, 성능과 신뢰성을 동시에 확보하는 설계가 요구된다.

그러나 생성형 AI의 도입은 실제와 다른 내용을 사실처럼 생성하는 환각(hallucination)과 자동화 편향(automation bias)으로 인한 산출물 신뢰성 문제를 수반한다[8]. 특히 감사와 같이 전문적 판단과 증빙 기반 추론이 요구되는 영역에서, 생성형 AI와 인간 전문가의 산출물

을 동일 조건에서 비교·검증한 실증 연구는 국내외적으로 제한적이다. 나아가 ITGC 평가조서라는 구체적 실무 단위에서 동일한 증빙 조건 하에 인간과 AI의 판정을 직접 대조한 연구는 거의 없으며, 생성형 AI의 환각을 통제된 조건에서 정량적으로 검증한 사례도 부족하다. 이러한 학술적 공백이 본 연구의 출발점이다.

본 연구는 ICFR 영역에서 생성형 AI를 활용한 ITGC 평가조서 작성 자동화 모델을 제안하고, 해당 모델의 실무 적용 가능성을 실증적으로 평가한다. 본 연구에서는 다음의 세 가지 연구문제를 설정하였다. (1) 구조화된 프롬프트와 자료형 안전 출력을 적용한 생성형 AI가 숙련 평가자와 동등한 수준의 ITGC 평가조서를 작성할 수 있는가? (2) 통제된 환경에서 생성형 AI의 환각은 어느 수준까지 억제되는가? (3) 실무자는 생성형 AI 기반 ITGC 자동화 시스템을 어떻게 수용하며 HITL은 어느 수준에서 요구되는가? 본 연구의 주요 기여는 다음과 같다.

첫째, ITGC 평가조서 작성이라는 구체적 실무 단위에서 생성형 AI와 숙련 평가자의 산출물을 동일한 증빙 조건에서 직접 대조한 실증 근거를 제시한다.

둘째, 8-Block 구조화 프롬프트와 자료형 안전 출력을 결합한 조서 자동화 설계를 제안하고, 사전에 정의한 조작적 기준에 따라 통제된 조건에서 환각 발생을 정량적으로 관찰·분석한다.

셋째, 성능 지표에 더하여 실무자 설문을 병행함으로써,

지표만으로는 포착되지 않는 실무 수용성과 인간 개입(HITL)에 대한 인식을 함께 규명한다.

본 논문의 구성은 다음과 같다. 제2장에서는 ITGC 평가와 생성형 AI 활용에 관한 배경과 선행연구를 정리한다. 제3장에서는 제안하는 조서 자동화 시스템의 구조와 8-Block 프롬프트, 자료형 안전 출력, 파라미터 제어 설계를 기술한다. 제4장에서는 실험 설계와 함께 효율성·정확성·환각 억제 및 실무자 수용성에 대한 분석 결과를 제시한다. 제5장에서는 결과의 함의와 인간-AI 협업 구조를 논의하고, 제6장에서 연구의 기여와 한계 및 향후 연구 방향을 제시한다.

II. Preliminaries

1. The Necessity of ITGC Assessment and Automation

IT 일반통제(ITGC)는 재무보고에 영향을 미치는 IT 시스템의 접근관리·변경관리·운영관리 등을 대상으로 하며 [9][10], 응용통제가 신뢰성 있게 작동하기 위한 기반 환경을 제공한다. ITGC의 운영 효과성 평가는 표본 추출, 증빙 수집, 증빙 대조, 조서 작성의 반복적 절차로 구성되어 상당한 투입 인력과 전문적 판단을 요구한다. 실제로 대형 회계법인의 조사에 따르면 기업의 SOX 준수에는 연평균 약 11,800시간이 투입되며 전체 통제의 약 80%가 수작업 또는 IT 의존적 수작업 통제로 수행되는 것으로 보고되는 등 [11], 수작업 중심 점검의 비효율은 선행연구에서도 지속적으로 지적되어 왔다 [8][9]. 특히 접근관리와 변경관리 영역은 권한 부여·회수, 프로그램 변경 승인 등 요청·승인·반영 간의 일관성을 다수의 증빙에 걸쳐 교차 확인해야 하므로, ITGC 중대결함이 상대적으로 빈번하게 보고되는 영역으로 알려져 있다. 기존의 수작업 기반 점검은 반복적 증빙 검토로 많은 시간과 비용이 소요되고 평가자의 숙련도에 따라 판단 편차가 발생할 수 있다 [12][13]. 이에 생성형 AI를 활용한 자동화는 업무 효율성 제고와 평가 일관성 확보의 대안으로 제시되고 있으며, 동일한 서비스 품질을 유지하면서 투입 원가를 절감할 수 있다는 점에서 감사 서비스 제공자의 경쟁력 확보에도 기여할 수 있다.

2. Application and Limitations of Generative AI in Auditing

생성형 AI를 회계·감사 과업에 적용한 선행연구들은 실무 활용 가능성과 운영 효율 향상을 실증적으로 검토해 왔다 [4][5]. Eulerich et al. [4]은 ChatGPT가 회계·감사 과업

전반에서 상당한 성과를 보이거나 복잡한 판단 과업에서는 성능 편차가 크다는 점을 실증하였고, Emmett et al. [5]은 내부감사 프로세스에 생성형 AI를 적용한 결과 감사 계획 수립과 문서 초안 작성 단계에서 실질적인 효율 향상이 나타남을 확인하였다. 경험 수준이 다른 감사인과 AI의 상호작용을 부정위험 평가 과업에서 실험적으로 분석한 연구에서는, 생성형 AI가 초보 감사인의 위험 식별 역량을 숙련자 수준 이상으로 견인하는 한편, 탐지된 위험을 실질적인 감사 계획으로 통합하는 고차원적 의사결정 영역에서는 지원 효과가 제한적임이 확인되었다 [14]. 이는 생성형 AI가 구조화된 식별 작업에서는 유효하나 최종 판단 단계에서는 여전히 인간의 역할이 필요함을 시사한다.

그러나 이러한 연구는 주로 기술적 성능의 시연이나 운영 효율에 초점을 두었으며, 대부분 일반적인 감사 업무에 한정되어 특정 통제 단위에서 산출물의 신뢰성을 실증한 사례는 부족하다. 한편 생성형 AI의 감사 적용에는 기술적·윤리적 위험이 수반된다. 실제 존재하지 않는 정보를 사실처럼 생성하는 환각은 통제의 결함을 은폐하거나 정상 통제를 결함으로 오판하게 할 수 있으며, 감사인이 AI 산출물에 과도하게 의존할 경우 독립적 판단과 전문적 의구심이 약화되는 자동화 편향이 발생할 수 있다 [15]. 이러한 위험은 전문 지식이 요구되는 감사 영역에서 특히 중요하게 작용한다. 또한 데이터 프라이버시, 알고리즘 편향, 책임 소재와 같은 윤리적 고려가 병행되어야 하며, 대형 회계법인의 도입 현장 분석에서도 단순 AI는 널리 활용되거나 복잡한 AI 활용은 투명성·편향·과잉 의존 우려로 인해 아직 개발·검증 단계에 있음이 보고되었다 [3]. 이는 생성형 AI를 감사에 도입하기 위해서는 산출물의 신뢰성과 검증 가능성을 확보하는 체계가 선결되어야 함을 의미하며, 본 연구가 구조화된 프롬프트와 자료형 안전 출력을 통해 다루고자 하는 지점이 바로 이 부분이다. 나아가 프롬프트의 구조화 수준이 대규모 언어모델(Large Language Model, LLM)의 추론 능력과 출력 품질에 유의한 영향을 미친다는 점도 보고되어, 프롬프트 설계가 성능의 핵심 변인임을 뒷받침한다 [16][17].

3. Distinctive Contributions of This Study

선행연구는 주로 감사 과업 전반에서 생성형 AI의 성과를 시연하거나 [4][5] 연구 동향을 개관하는 [6] 데 초점을 두었으며, 인간과 AI의 성과 비교 역시 특정 과업이나 제한된 맥락에 국한된 경우가 많았다. 그 결과 ITGC 평가조서 작성이라는 구체적 실무 단위에서, 동일한 증빙 조건 하에 인간과 AI의 판정을 직접 대조한 실증 분석은 충분히 축적되

지 않았다. 본 연구는 다음 네 가지 측면에서 선행연구와 차별화된다. 첫째, 분석 단위 측면에서 기존 연구가 감사 업무 전반이나 특정 과업을 대상으로 한 것과 달리, 본 연구는 ITGC 평가조서 작성이라는 구체적 실무 단위를 분석 대상으로 구체화하였다. 둘째, 비교 방식 측면에서 기존 연구가 AI의 성능을 단독으로 측정하거나 제한된 맥락에서 비교한 것과 달리, 본 연구는 동일한 증빙 조건에서 생성형 AI와 숙련 평가자의 산출물을 직접 대조하였다. 셋째, 신뢰성 검증 측면에서 기존 연구가 환각 위험을 주로 개념적으로 논의한 것과 달리, 본 연구는 구조화 프롬프트와 자료형 안전 출력이 결합된 조건에서 환각 발생을 정량적으로 분석하였다. 넷째, 수용성 측면에서 본 연구는 성능 지표에 더하여 실무자 설문을 병행함으로써, 지표만으로는 포착되지 않는 실무 도입 조건을 규명하였다. 따라서 본 연구는 생성형 AI 기반 ITGC 자동화의 성능과 실무 수용성을 동일한 실험 환경에서 함께 검증함으로써 관련 분야의 실증적 연구 기반을 확대하고, 후속 연구에 활용 가능한 검증 프레임워크를 제안한다. 이러한 접근은 생성형 AI 기반 ITGC 자동화를 체계적으로 검증할 수 있는 연구 틀을 마련하였다는 점에서 학술적 의의를 가진다.

III. The Proposed Scheme

1. System Architecture

ITGC 운영 효과성 평가는 일반적으로 (1) 모집단 추출, (2) 표본 추출, (3) 증빙 수집, (4) 증빙 대조·판정, (5) 조서 작성의 5단계로 구성된다. 이 중 모집단 추출과 증빙 수집은 피감사 기업의 IT 담당자가 수행하며, 나머지 평가 업무는 평가 수행 주체가 담당한다. 본 연구의 생성형 AI 시스템은 이 가운데 판단 부하가 크고 반복적인 (4) 증빙 대조·판정과 (5) 조서 작성을 자동화 대상으로 한정한다. 즉 평가자가 사전에 등록한 모집단 조직도와 업로드된 증빙을 기반으로, 시스템이 통제 유효성을 판정하고 엑셀 조서를 산출한다.

제안 시스템은 클라이언트 계층(Client Layer), 애플리케이션 계층(Application Layer), 데이터 계층(Data Layer), 외부 연동 계층(External Layer)의 4계층으로 구성된다(Fig. 1).

데이터의 전달과 제어 흐름(Call Flow)은 시스템 복잡도를 낮추기 위해 ‘클라이언트→애플리케이션→데이터’ 및 ‘애플리케이션→외부 연동’ 방향의 단방향(unidirectional)으로 엄격히 통제하였다. 이러한 단방향

설계는 세 가지 이점을 위한 것이다. 첫째, 계층 간 순환 참조를 배제하여 데이터 흐름을 예측 가능하게 하고 처리 경로의 소프트웨어적 재현성을 높인다. 다만 이는 파이프라인 수준의 재현성을 의미하며, 생성형 AI 모델 출력 자체의 결정성을 보장하는 것은 아니다. 둘째, 계층 간 결합도를 낮추어 유지보수와 확장을 용이하게 한다. 셋째, 판정에 사용된 증빙과 처리 경로를 일관되게 역추적할 수 있어, 감사에서 요구되는 추적 가능성(traceability)을 뒷받침한다. 클라이언트 계층은 통제된 실험 환경을 위한 UI를 제공하고, 애플리케이션 계층은 비동기 처리에 최적화된 서버에서 증빙 분석 파이프라인을 수행한다. 데이터 계층은 분석 입력을 가공 없이 스냅샷으로 직렬화하는 구조를 채택하여 중단된 분석의 재개를 지원하고, 판정 결과와 산출 조서를 구조화된 형태로 보존하여 추적 가능성을 확보하며, 외부 연동 계층은 생성형 AI(Anthropic Claude) 호출과 파라미터 제어를 담당한다.

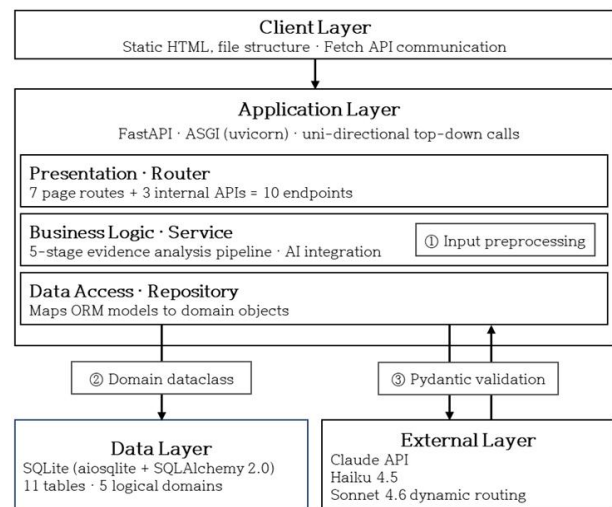


Fig. 1. Four-Layer Architecture Diagram

2. 8-Block Structured Prompt

프롬프트의 구조화 수준이 LLM의 추론 능력과 출력 품질에 유의한 영향을 미친다는 점은 선행연구에서 지속적으로 보고되어 왔으며[16][17], 특히 복잡한 과업을 하위 구성요소로 분해하여 제시하는 구조화 기법이 모델의 추론 정확도를 높이는 것으로 알려져 있다. 본 연구의 8-Block 구조는 이러한 선행연구의 원리를 ITGC 평가 맥락에 적용한 것으로, 기존 프롬프트 기법이 지시를 단일 텍스트로 전달하는 데 비해, 본 연구가 설계한 8-Block 구조는 실제 평가자의 분석 절차를 반영하여 분석에 필요한 정보를 역할별 블록으로 분리하고 이를 누적적으로 조합한다. 구체적으로, (1) 분석 컨텍스트(Analysis Context)

는 평가 대상 시스템의 환경과 통제 목적을 정의하고, (8) 추출 프롬프트는 모든 통제에 공통으로 적용되는 추출 항목을 규정하는 2개의 필수 블록이다. 여기에 통제의 특성에 따라 (2) 조직도 맥락, (3) 모집단 맥락, (4) 권한 사전, (5) 평가 기간 범위, (6) 세부 수행 지시사항, (7) 출력 형식 지시의 최대 6개 선택적 블록이 조합되며, 실제 삽입 순서는 Algorithm 1과 같다.

Algorithm 1. 8-Block Prompt Composition Procedure

```

Input: evidence, contexts, options
Output: composed prompt
blocks ← <> // fixed order below; append only blocks whose
condition holds
① Analysis Context (Required)
  append([Analysis Context]: control activity-description,
  evidence type, sample no.)
② Organization Chart Context (Optional)
  IF contexts.org_chart ≠ ∅ THEN append([Org Chart
  Context]: org summary)
③ Population Context (Optional) - two forms by purpose
  IF contexts.population ≠ ∅ THEN
    IF is_population_validation THEN append([Raw Data]:
    population source)
    ELSE append([Population Context]: full population)
④ Role Dictionary (Optional)
  IF contexts.role_list ≠ ∅ THEN append([Role Dictionary]:
  role-code definitions)
⑤ Evaluation Period Coverage (Optional)
  IF audit_period is set THEN append([Evaluation Period
  Coverage])
⑥ Detailed Instructions (Optional)
  IF evidence.instructions ≠ ∅ THEN
    append([Detailed Instructions]: judgment
    criteria-review-comment template)
⑦ Output Format Instruction (Optional) - by column
  configuration
  IF evidence.extra_columns ≠ ∅ THEN
    IF table-type THEN append([Table Output Instruction]:
    judgment-data column rules)
    ELSE append([Required Extraction Items]: column list)
⑧ Extraction Prompt (Required) - one of three
  IF is_population_validation THEN append(Population
  Validation Prompt)
  ELSE IF custom_prompt ≠ ∅ THEN
    append(Company-specific Custom Prompt)
  ELSE append([Common Extraction Target] Default Prompt)
RETURN join(blocks)

```

이러한 모듈화는 여러 가지 이점을 가지며, 요약하면 첫째, 모델이 각 판단 근거를 명시적으로 참조하도록 유도하여 비정형 증빙에 대한 추론의 일관성을 높이는 것을 목적으로 한다. 둘째, 모든 통제에 동일한 프롬프트를 적용하는 방식과 달리 통제 특성에 따라 필요한 블록만 선택적으로 조합할 수 있다. 셋째, 블록 단위의 재사용과 유지보수가 용이하여 통제 유형이 확장되더라도 프롬프트 설계 비

용이 크게 증가하지 않는다. 8-Block의 구성과 각 블록의 역할은 Table 1과 같다.

Table 1. Structure of the 8-Block Prompt

Type	Block	Role
Required	Analysis Context	Defines target system environment and control objectives
Required	Extraction Prompt	Specifies extraction items common to all controls
Optional	Organization Chart Context	Provides organizational structure and personnel data
Optional	Population Context	Provides target population data
Optional	Role Dictionary	Provides role definitions
Optional	Evaluation Period	Specifies the evaluation period
Optional	Detailed Instructions	Provides control-specific judgment guidelines
Optional	Output Format Instruction	Specifies table columns or required extraction items

각 통제의 평가는 필수 블록에 통제 특성에 맞는 선택 블록을 결합하여 프롬프트를 구성하고, 이를 라우팅된 모델에 전달하여 구조화된 판정 결과를 반환받는 절차로 수행된다. 재현에 필요한 조합 및 판정 절차는 다음 의사코드로 제시한다. 아울러 한 통제를 대상으로 시스템 명칭을 치환한 조립 결과를 Fig. 2에 발췌하여 보이고, 같은 통제의 판단 규칙 전문과 입력 증빙 및 출력 예시는 Fig. 4~6에 제시한다.

분석 컨텍스트 블록은 통제명 등의 값 중 하나라도 존재하면 삽입되나 운영상 통제명이 항상 존재하므로 사실상 필수로 작동한다. 모집단-출력 형식-추출 프롬프트 블록은 통제 목적이나 컬럼 구성에 따라 내용이 변형되며, 추출 프롬프트는 기본 프롬프트를 기준으로 하되 회사별 커스텀 프롬프트가 등록된 경우 이를 대체한다. 교차 검증이 요구되는 통제는 역할별 증빙 추출(1단계)과 조인 키 기준 통합 판정(2단계)으로 확장된다.

Fig. 2는 교차 검증이 요구되는 통제의 2단계(통합 판정) 프롬프트이다. 이 경로에서는 모집단 컨텍스트 자리에 1단계에서 역할별로 추출한 결과가 삽입되며, 권한 사전이 그보다 앞서 배치된다. 평가기간 커버리지 블록(⑤)은 해당 통제에 평가기간이 설정되지 않아 삽입되지 않았다.

[분석 컨텍스트]

통제 활동: Application 계정/권한 회수

통제 설명: 퇴사자 및 부서이동자 발생 시 ... 퇴사자 계정은 발령일을 포함하여 즉시 회수하고, 부서이동자의 경우 인수인계 사유로 발령 후 5일 이내에 권한을 회수한다.

처리 모드: 교차 검증 - 최종 통합 분석

처리 대상 발령자_사번: 20건

[조직도 컨텍스트] / [조직도]

SOD(직무 분리) 판단 시 아래 조직도를 참고하세요.

회사명: HRM 시스템 ... (부서·직급·인원 1,402자 생략)

[권한 사전]

권한코드 | 권한명 | 설명

ROLE_USR | 일반사용자 | 기본 권한

ROLE_HR_OP | HR운영자 | 인사 시스템 운영자 (인사운영팀)

ROLE_ACC_OP | 계정/권한 관리자 | 모든 계정·권한의 운영 관리자 (ICT운영팀)

... ('주요 권한' 열과 나머지 행 생략, 768자)

[증빙별 추출 데이터]

[인사발령]

No | 발령구분 | 발령자_사번 | 발령자_성명 | 소속부서 | 발령일 | ...

1 | 부서이동 | 2503005 | 고우리 | 인사기획팀 | 2025-12-30 | ...

2 | 퇴직 | 2312001 | 정서우 | 재경팀 | 2025-12-29 | ...

... (1,558자 생략)

[권한회수이력]

No | 처리유형 | 회수_사번 | 성명 | 권한명 | 권한코드 | 처리일시

1 | 계정회수 | 2312001 | 정서우 | - | - | 2025-12-29

2 | 권한회수 | 2312001 | 정서우 | 일반사용자 | ROLE_USR | 2025-12-29

... (683자 생략)

[인사발령 세부 지시사항]

[추출 대상 필터]

발령구분이 '퇴사', '퇴직' 또는 '부서이동'인 행만 추출 대상이다.

[행 단위 빈칸 보존 규칙 - 매우 중요]

- 추출 대상 표의 각 행은 독립이다. 한 행의 빈 셀을 다른 행의 값으로 채우지 말 것.

[교차검증 2단계 케이스 분류 및 Att 판정 기준]

- 케이스 A (퇴직 + 기본권한): 계정회수일 <= 발령일이면 Pass

- 케이스 C (부서이동 + 기본권한): 회수 대상 없음. 자동 Pass

... (740자 생략)

[검토의견 메시지]

Pass (케이스 A): "[사번]/[성명]([현부서]/[발령직급])은/는 퇴사 발령([발령일])에 따라 계정이 적시에 회수(계정: [계정회수일])되었음을 확인함"

[출력 컬럼 순서]

발령자_사번, 발령자_성명, 발령구분, 발령일, ..., 검토의견, Att 1

[교차 검증 지시] / [출력 규칙]

위 데이터를 '발령자_사번' 기준으로 매핑하여 교차 검증하세요.

1. 발령자_사번마다 하나의 ControlRow를 rows에 생성하세요.

3. Att 1 값은 'Pass' / 'Fail' / '확인필요' 중 하나.

Fig. 2. Excerpt from the Assembled 8-Block Prompt (Application Account and Access Revocation)

3. Type-Safe Output

출력의 구조적 일관성을 강제하고 형식 오류를 사전에 차단하기 위해, 런타임에서 자료형과 제약조건을 검증하는 Pydantic 기반 자료형 안전 출력(Type-Safe Output)을 채택하였다(Fig. 3).

```
class JudgmentResult(StrEnum):
    PASS = "PASS"
    FAIL = "FAIL"
    UNCERTAIN = "확인필요"
class ITGCExtractionResult(BaseModel):
    document_type: str
    document_date: date | None = None
    extraction_confidence: float = Field(ge=0.0, le=1.0)
    overall_result: JudgmentResult
    rows: list[ControlRow]
    result: ITGCExtractionResult =
client.chat.completions.create(
    model=_model, temperature=0.0,
    response_model=ITGCExtractionResult,
    messages=[{"role": "user", "content": [content_block,
    prompt_block]}],
```

Fig. 3. Example Code for Pydantic Schema Validation

위 스키마는 결론값을 세 가지 열거형으로 제한하고 추출 신뢰도를 [0,1] 범위로 강제하며, 확인 불가능한 필드에 대해 임의 추정 대신 null을 반환하도록 제약한다.

전통적인 자연어 출력 방식은 형식이 불안정하고 후속 처리 과정에서 파싱(parsing) 오류가 발생하기 쉬운 반면, 자료형 안전 출력은 사전에 정의된 스키마에 부합하지 않는 응답을 형식 검증 단계에서 배제한다. 구체적으로, 통제 결론값은 세 가지 열거형(PASS, FAIL, 확인필요)으로만 제한되어 모델이 정의되지 않은 판정을 생성할 수 없으며, 추출 신뢰도는 0.0~1.0 범위의 실수로 강제되고, 스키마를 위반한 출력은 즉시 폐기된다. 본 시스템은 위반 시 재생성을 시도하지 않고 오류로 처리하며, 재시도는 API Rate Limit (HTTP 429)에 한해 적용된다. 특히 판단 근거가 부족한 경우 모델이 임의로 결론을 내리는 대신 '확인필요' 상태로 분류하도록 설계하여, 불확실한 사안을 인간 검토로 이관하는 안전장치를 마련하였다. 이를 통해 정의되지 않은 형식이나 범위를 벗어난 값, 열거형에 없는 판정이 최종 산출물에 유입되는 것을 형식 수준에서 걸러낸다. 다만 자료형 안전 출력은 출력의 형식과 값의 범위를 검증하는 장치로서, 추출된 값이 실제 증빙과 내용상 일치하는지까지 보장하지는 않는다. 즉 형식 오류의 차단과 의

미적 환각의 방지는 구분되며, 후자는 별도의 증빙 대조 절차를 통해 검증되어야 한다. 본 실험 조건에서 스키마 검증을 통과하지 못한 비정형 응답은 파서 단계에서 폐기되고 오류로 처리되었으며, 그 결과 전체 154건의 분석에서 형식 오류가 최종 조서에 유입된 사례나 환각으로 판정된 사례는 관찰되지 않았다.

한편 검토의견 문장은 모델이 자유롭게 생성하도록 두지 않고, 후처리 단계에서 코드가 다시 작성한다. 모델이 통합행을 구성할 때 검토의견 텍스트 안에 다른 행의 날짜·식별자·성명을 잘못 인용하는 오류가 관찰되었기 때문이다. 통제별 문구 템플릿을 코드에 보유하고, 해당 행이 보유한 필드 값만으로 자리표시자를 치환하여 검토의견을 생성한 뒤 모델의 출력을 이 결과로 대체한다. 따라서 검토의견에 인용되는 모든 값은 정의상 같은 행의 필드와 일치하며, 행 간 값 혼입이 최종 조서에 유입될 경로가 제거된다. 프롬프트에도 동일한 문구 템플릿이 포함되어 있으나 이는 모델이 판정 결과를 일관된 형식으로 기술하도록 유도하기 위한 것이며, 최종 산출물의 문장은 위 후처리 결과이다.

Fig. 4는 Fig. 2와 같은 통제에 적용되는 판단 규칙의 전문이다. 발령구분과 권한 유형의 조합으로 케이스를 정의하고, 케이스별로 회수 대상과 기한을 달리 적용한다. 판정 결과의 서술 문구까지 템플릿으로 고정하여, 동일한 케이스에 대해 동일한 형식의 검토의견이 산출되도록 하였다. 케이스 B와 D의 검토의견 문구는 케이스 A와 같은 도입부를 생략하여 표기하였으나, 판정 조건은 생략하지 않았다. 다른 통제의 판단 규칙도 같은 구조를 따르며, 케이스 정의와 기한만 통제별로 달라진다.

[교차검증 2단계 케이스 분류 및 Att 판정 기준]
아래 기준은 교차검증 2단계에서 [계정/권한 회수 이력]과 사번 기준으로 매핑한 후 적용한다.
권한명은 [계정/권한 회수 이력]에서 가져오며, 발령구분은 이 데이터에서 가져온다.
케이스 분류 (발령구분 × 권한명):
- 케이스 A (퇴직/퇴사 + 기본권한): 발령구분 = '퇴사' OR '퇴직' AND 권한명 = '일반사용자' → 계정만 회수 대상. 계정회수일 <= 퇴사(퇴직)발령일이면 Pass.
- 케이스 B (퇴직/퇴사 + 특수권한): 발령구분 = '퇴사' OR '퇴직' AND 권한명 ≠ '일반사용자' → 계정 + 권한 동시 회수 대상. 계정회수일 AND 권한회수일 <= 발령일이면 Pass.

Fig. 5는 비정형 증빙에서 조서 항목을 추출하는 1단계 판독의 예시로, 프로그램 변경 요청 승인 통제의 변경 요청 승인 문서이다. 계정·권한 회수 통제의 증빙은 인사발령 이력과 권한 회수 이력의 두 구조화 테이블로만 구성되

- 케이스 C (부서이동 + 기본권한): 발령구분='부서이동' AND 권한명='일반사용자' → 회수 대상 없음. 자동 Pass.
 - 케이스 D (부서이동 + 특수권한): 발령구분='부서이동' AND 권한명≠'일반사용자' → 권한만 회수 대상. 권한회수일 <= 발령일 + 5일이면 Pass.
 Att 1. 적시 회수 여부
 - 케이스 A: 계정회수일 <= 발령일이면 Pass. 이력 없거나 초과 시 Fail.
 - 케이스 B: 계정회수일 <= 발령일 AND 권한회수일 <= 발령일이면 Pass. 하나라도 미충족 또는 이력 없으면 Fail.
 - 케이스 C: 자동 Pass (회수 대상 없음).
 - 케이스 D: 권한회수일 <= 발령일+5일이면 Pass. 이력 없거나 초과 시 Fail.
 [중요 - 사전 회수 처리]
 회수일이 발령일보다 빠른 경우(사전 회수)도 통제 위반이 아니며 Pass로 판정한다.
 Fail은 오직 회수일이 발령일(케이스 D는 발령일+5일)을 '초과'한 경우에만 적용한다.
 회수 이력이 아예 없는 경우만 Fail/확인필요로 분류한다.
 [판정 기준 공통]
 - Pass: 해당 케이스의 모든 조건 충족
 - Fail: 조건 미충족 사항이 하나라도 식별된 경우
 - 확인필요: 회수 이력 데이터 부재, 증빙 누락, 데이터 형식 오류로 판단 불가 시
 [검토의견 메시지]
 Pass (케이스 A): "[사번]/[성명]([현부서]/[발령직급])은/는 퇴사 발령([발령일])에 따라 계정이 적시에 회수(계정: [계정회수일])되었음을 확인함"
 Pass (케이스 B): "... 계정 및 권한([권한명])이 적시에 회수(계정: [계정회수일]/권한: [권한회수일])되었음을 확인함"
 Pass (케이스 C): "... 부서이동 발령([발령일])에 따른 기본권한(일반사용자) 보유자로 회수 대상 권한이 없음을 확인함"
 Pass (케이스 D): "... 부서이동 발령([발령일])에 따라 권한([권한명])이 적시에 회수([권한회수일])되었음을 확인함"
 Fail 공통: "[사번]/[성명]의 [퇴사/퇴직/부서이동] 발령([발령일])에 따른 계정 또는 권한([권한명]) 회수가 기준 내 처리되지 않았음 (계정: [계정회수일]/권한: [권한회수일])"

Fig. 4. Decision Rules for Application Account and Access Revocation Controls: Original Prompt Used in the Initial Experiment and the Pre-Refinement Rerun

주: 본 그림은 보완 이전 실험에 사용한 프롬프트 원문을 보존한 것이다. 회수 이력 부재에 대한 FAIL과 확인필요 규칙의 적용 우선순위가 명확하지 않았으며, 보완 조건에서는 대상 시스템의 처리 전제와 판정 규칙의 우선순위를 추가하였다.

전자결재 · 프로그램 변경 요청서

출근기록 기반 재직 확인 기능 개선

요청부서: 요청자: 요청일:
 인사운영팀 노민석 (대리) 2025-01-08
 문서번호: CHG-2025-111 | 배포번호: REL-2025-003 | 테스트 문서 번호: TS-2025-002

1. 변경 기본 정보

업무 명칭: 변경 유형:

변경 우선순위: 요청 부서 / 요청자:

업무 목적 (BUSINESS PURPOSE):

기능 요구사항 (FUNCTIONAL REQUIREMENT)
 ① ATD 관련 모듈 및 DB 정합성 검증 ② 출근기록 기반 재직 확인 기능 개선 반영 및 예외 로직 처리

변경 상세 설명

근태관리 시스템의 관련 모듈을 수정하여 '출근기록 기반 재직 확인 기능 개선' 처리함. 영향 범위: 핵심 처리 모듈, 화면 출력 모듈, 데이터 처리 배치. 변경 대상 파일: 총 5개 (Java 3개, SQL 2개)

처리결과: ☒ 완료 (COMPLETED) 배포 완료일: 2025-01-31

2. 전자결재 승인선

승인 흐름: 현업 사용자 → 현업 부서장 → 릴리즈 담당자

현업 사용자 (기안)	현업 부서장 (1차 승인)	릴리즈 담당자 (최종 승인)
<div>● 승인됨</div> <div>노민석</div> <div>인사운영팀 / 담당</div> <div>2025-01-08</div>	<div>● 승인됨</div> <div>성미나</div> <div>인사운영팀</div> <div>2025-01-08</div>	<div>● 승인됨</div> <div>표민수</div> <div>ICT운영팀</div> <div>2025-01-31</div>

Fig. 5. Unstructured Audit Evidence (Change Request Approval Document, CHG-2025-111)

어 비정형 증빙이 없으므로, 입력 증빙의 예시는 이 통제에서 택하였다. 문서는 서술형 본문과 전자결재 승인선이 혼재된 형식이며, 승인 정보는 표가 아닌, 결재 흐름의 시각적 배치로만 구분된다. 추출은 승인선의 순서에서 1차 승인자를 식별하며, 기안자와 최종 승인자를 1차 승인자 자리로 옮기지 않는다.

[케이스 C(Pass), 케이스 A(Fail)]

필드	Pass · 부서이동 (케이스 C)	Fail · 퇴직 (적시 회수 미충족)
발령자_사번	2301003	2512003
발령자_성명	기은주	김지수
발령구분	부서이동	퇴직
발령일	2025-11-24	2025-11-19
발령직급	차장	-
전 부서	장기상품팀	-
현 부서	현장심사팀	-
회수_사번	미확인	2512003
성명	미확인	김지수
권한명	미확인	-
권한코드	미확인	-
계정상태	미확인	회수
계정회수일	null	2025-11-21
권한회수일	null	2025-11-21
Att 1	Pass	Fail
overall_result	PASS	FAIL

검토의견	2301003/기은주 (현장심사팀, 차장) 은/는 부사이동 발령 (2025-11-24)에 따라 기본 권한(user) 을 보유하고 있어 권한 회수 대상자가 아님	[Fail] 김지수(2512003)의 퇴직 발령 (2025-11-19)에 대한 계정회수일 (2025-11-21)이 발령일 이후에 수행되어 적시에 회수되지 않았으므로 조건을 충족하지 못함
------	---	---

Fig. 6. Example of a Generated Audit Working Paper Row

각 행은 조서 컬럼 16개의 값과 속성별 판정(Att 1), 종합 판정(overall_result), 검토의견을 보여준다(Fig. 6). 두 건은 Fig. 4의 케이스 C와 케이스 A에 각각 해당하며, 이 Pass 사례는 회수 이력에 대응 레코드가 없어 권한명이 '미확인'으로 추출된 상태에서 기본권한 보유자로 해석된 원실험의 산출물이다. 회수 이력 부재만으로 기본권한 보유 여부를 확정할 수 없으므로, 본 사례는 판정 기준의 공백과 재실행 변동성을 설명하는 예시로 제시한다. 검토의견은 앞서 기술한 후처리 단계에서 코드가 보유한 문구 템플릿으로 다시 작성된 결과이므로 Fig. 4의 프롬프트 내 문구와 표현이 일부 다르며, 인용된 날짜와 식별자는 모두 같은 행의 필드 값이다.

4. Parameter Control and Model Routing

결과 산출의 변동성을 줄이기 위해 생성 온도(temperature)를 0.0으로 설정하고 top-p는 별도로 지정하지 않아, 확률적 샘플링이 출력에 미치는 영향을 최소화하였다. 이는 출력의 변동성을 낮추기 위한 조치이나, API 기반 생성형 AI는 temperature를 0으로 설정하더라도 서버 측 구현, 모델 갱신, 부동소수점 연산, 동률 토큰 처리 등으로 인해 완전한 결정성이 보장되지는 않는다. 따라서 본 연구의 파라미터 설정은 출력 안정성을 높이기 위한 조치로 이해되어야 하며, 모델 수준의 완전한 재현성을 의미하지는 않는다. 또한 작업의 추론 난이도에 따라 모델을 동적으로 할당하는 라우팅 구조를 적용하였다. 단순 단일 추출이나 모집단 완전성 검증과 같이 부하가 낮은 작업에는 경량 모델(Claude Haiku 4.5)을, 다수 증빙의 교차 검증·자동 분류·다중 파일 분석과 같이 복잡도가 높은 작업에는 고성능 모델(Claude Sonnet 4.6)을 할당함으로써, 판정 정확성과 연산 비용의 균형을 도모하였다. 재현성 확보를 위해 실험에 사용한 API 구성을 다음과 같이 명시한다. 모델 호출은 Anthropic Claude API를 기반으로 하였으며, 애플리케이션과 API 간 통신 구간에는 instructor 라이브러리를 결합하여 Claude의 도구 호출(Tool-use)

기능을 통해 사전 정의된 Pydantic 스키마 내에서만 결과를 반환하도록 강제하였다. 모델 배정은 결정론적 라우팅 함수를 통해 이루어지며, 동일한 조건의 입력에 대해서는 항상 동일한 모델이 대응되도록 설계하였다. 작업 유형에 따라 최대 생성 토큰(max_tokens)을 동적으로 할당하였고, API Rate Limit(HTTP 429)에 한해 재시도를 적용하고, 그 밖의 오류와 재시도로 해소되지 않는 오류에 대해서는 예외를 발생시켜 연산을 중단하였다. 재현성을 위해 실험 환경을 명시하면, 판정 경로에는 Claude Haiku 4.5(claude-haiku-4-5-20251001)와 Claude Sonnet 4.6(claude-sonnet-4-6) 두 모델을 사용하였으며(전자는 날짜가 지정된 스냅샷 식별자이고 후자는 날짜가 없는 별칭으로, 후자는 호출 시점에 적용된 스냅샷을 응답에서 식별할 수 없다), Python 3.12.3 환경에서 anthropic 0.96.0·instructor 1.14.5·pydantic 2.13.3 라이브러리를 기반으로 2026년 5월에 실험을 수행하였다. 호출에는 정식(GA) 엔드포인트만 사용하였고 베타 기능은 사용하지 않았으며, 요청 timeout은 별도로 설정하지 않아 SDK 기본값을 따랐다. 다만 API 기반 생성형 AI의 사양은 버전에 따라 변동될 수 있으므로, 본 결과는 해당 시점의 모델을 기준으로 한다. 출력 변동성을 낮추기 위해 temperature=0.0으로 고정하였으며, 별도의 반복 호출이나 self-consistency 기반 다수결 절차는 사용하지 않았다. 재시도는 API 호출이 Rate Limit (HTTP 429)에 해당하는 경우에 한해 지수 백오프(exponential backoff) 방식으로 이루어지며, 최초 호출을 포함해 총 3회까지 시도한다. 판정 자체를 여러 번 생성하여 다수결로 확정하는 구조는 아니다.

IV. Experiments and Results

1. Experimental Design and Data

본 연구는 접근관리·변경관리 영역의 6개 ITGC 통제를 대상으로, 동일한 증빙 데이터 세트에 대해 구조화된 생성형 AI와 숙련 평가자가 각각 평가조서를 작성하고 그 판정을 대조하는 대응 비교(paired benchmark) 방식으로 수행하였다. 이는 복수의 참여자를 무작위로 배정한 통제 실험이 아니라, 단일 숙련 평가자의 판정과 시스템 산출을 동일 증빙 조건에서 짝지어 비교한 설계이다. 대상 통제는 접근관리 영역의 계정·권한 요청 승인, 계정·권한 회수, 비밀번호 설정, 특수 권한 제한 4개와, 변경관리 영역의 개발·운영환경 분리, 프로그램 변경 요청 승인 2개로 구성하

였다[9][10]. 이들은 요청·승인·반영 간 일관성 확인이나 파라미터 검증, 권한 오남용 방지 등 실무에서 중대결함이 상대적으로 빈번하게 보고되는 대표 통제에 해당한다. 대조군인 인간 평가자는 K-SOX 기반 ITGC 평가 경험을 보유한 숙련된 전문가 1인을 평가자로 선정하였다. 해당 평가자는 본 연구의 저자로서, 시스템 개발자이자 평가자를 겸함에 따라 잠재적 이해상충이 존재할 수 있으며, 이는 본 연구의 중요한 한계로 인식한다. 다만 평가는 AI 시스템의 산출물을 참조하지 않고 독립적으로 수행되었다. 이 평가자에게는 생성형 AI 시스템과 동일한 증빙 세트와 통제 명세서가 제공되었으며, 외부 참조가 차단된 독립적 환경에서 실무의 1차 검토자(first-pass reviewer) 수준을 기준으로 평가조서를 직접 작성하였다. 해당 평가자가 작성한 평가조서를 판정의 기준값(reference)으로 설정하였다. 이는 숙련 평가자의 판단을 사실상의 정답으로 간주하는 설계로, 본 연구의 정확도는 생성형 AI의 판정이 숙련 평가자의 판정과 일치하는 정도로 정의된다. 평가 조건과 데이터 생성 절차는 다음과 같다. 실험에 사용한 가상 증빙은 평가 시스템과 별개의 프로젝트로 개발한 가상 HR 시스템에서 산출하였다. 해당 시스템의 화면과 문서 양식은 연구자가 제시한 구성에 따라 구현하였고, 시스템과 데이터 생성 코드는 Anthropic의 Claude Code 환경에서 Claude Opus 4.6·Sonnet 4.6·Haiku 4.5로 구성된 에이전트 팀이 작성하였다(2026년 4월). 정상·긴급·미승인 등 시나리오의 대상 사례와 조건값은 생성 과정에서 생성형 AI가 정하였으며, 연구자가 특정 사례를 결함으로 지정하지 않았다. 연구자는 정상 변경은 승인 후 배포하고 긴급 변경은 선포 후 사후 승인한다는 등의 업무 처리 원칙을 생성형 AI에 제시하였을 뿐, 개별 레코드의 내용이나 결함 해당 여부는 검토하지 않았다. 숙련 평가자는 각 사례의 결함 여부나 정상·결함의 분포를 사전에 알지 못한 상태에서 증빙만을 근거로 순차적으로 판정하였다. 또한 인간 평가가 먼저 수행되고 생성형 AI 평가가 이후에 이루어졌으므로, 두 주체는 서로의 판정 결과를 참조하지 않고 독립적으로 평가하였다. 다만 단일 평가자의 판정을 기준값으로 사용한 점은 여전히 타당도의 한계이므로, 이를 보완하기 위해 교차 검증이 요구되는 3개 핵심 통제(프로그램 변경 요청 승인, 계정·권한 요청 승인, 계정·권한 회수)에 대해 각 20건씩 총 60건을 선정하고, 원 평가자를 포함한 IT 감사 실무자 3인이 각각 독립적으로 평가조서를 작성하여 평가자 간 일치도를 검토하였다. 평가자 A는 CISA(Certified Information Systems Auditor) 시험에 합격

하였으며(자격 인증 요건인 실무경력 충족 절차는 진행 중), 정보시스템 감사 관련 실무를 3년째 수행하고 있다. 평가자 B는 ITGC 실무 경력 5년 및 정보보안 실무 경력 5년을 보유하고 있다. 세 평가자는 동일 법인 및 동일 부서에 소속되어 있어 완전히 독립적인 제3자 검증과는 성격이 다르나, 판단 기준은 개별 훈련이 아닌 글로벌 감사방법론(global audit methodology)에 따라 규정되고 개별 사안에 대한 최종 판단은 각 평가자가 독립적으로 수행한다는 점에서 일정 수준의 판단 독립성이 확보된다. 이는 완전한 독립적 검증이라기보다 조직 내 신뢰도 점검(within-organization reliability check)의 성격에 가까우며, 이러한 조직적 근접성이 평가자 간 일치도에 미쳤을 영향을 배제할 수 없음을 한계로 명시한다.

평가 대상 증빙은 실제 회계감사 현장의 ITGC 평가 환경을 정밀하게 모사하여 설계한 가상 데이터 세트로, 6개 통제 영역 전반에 걸쳐 유기적으로 연결된 증빙으로 구성하였다. 증빙에 등장하는 성명·사번·부서 등은 모두 가상으로 생성한 값으로 실존 인물이나 조직의 정보를 포함하지 않으며, 따라서 본문의 그림에는 이를 별도의 가림 처리 없이 수록하였다. 여러 증빙 간 교차 검증이 요구되는 3개 핵심 통제(프로그램 변경 요청 승인, 계정·권한 요청 승인, 권한 회수)에 대해서는 각각 40건 이상의 표본을 확보하였고, 증빙 구조가 단조로운 나머지 3개 통제는 실무 관행에 부합하는 수준으로 표본을 구성하였다. 결함 시나리오는 승인 누락 권한 부여, 퇴직자 계정 회수 지연, 배포 승인 누락 등 실제 업무에서 빈번한 핵심 위험과 통제 이탈 사례를 반영하여 설계하였으며, 정상 사례와 결함 사례가 특정 순서나 위치에 치우치지 않도록 혼재되도록 구성하여 평가 주체가 사전 확률에 의존하지 않고 각 증빙을 독립적으로 판단하도록 통제하였다. 이러한 절차를 통해 총 154건의 판정 단위를 확보하였으며, 두 주체의 판정 결과를 오차행렬(confusion matrix)로 대조하여 전체 정확도(accuracy)와 함께 민감도(sensitivity), 특이도(specificity), 정밀도(precision), F1 점수를 산출하였다. 이 중 판단 부하가 큰 계정·권한 요청 승인과 프로그램 변경 요청 승인 통제에 대해서는 오류 패턴을 추가로 분석하였다. 한편 판단 규칙의 출처는 감사 실무의 IT 일반통제 위험통제매트릭스(Risk Control Matrix)로, 증빙 데이터 생성 이전에 확정되었으며 특정 사례의 식별자를 지목한 예외 규칙은 포함하지 않는다. 또한 모집단 형태의 증빙(배포 이력, 인사발령, 권한 이력, 권한 회수 이력)의 값은 생성형 AI가 아니라 결정론적 파서가 읽는다. 다만 개발용

과 평가용 증빙을 분리하지 않아, 판단 규칙을 보완하는 과정에서 평가 대상 증빙의 실행 결과를 참조하였다. 보완은 긴급 변경·사후 승인에 대한 판정 분기 추가와 추출 오류 수정에 한정되었다. 그러나 평가 증빙의 실행 결과를 보완 과정에서 참조하였으므로 개발과 평가의 독립성은 확보되지 않았다. PASS 판정 건수의 유지 여부만으로 평가 자료에 대한 적합 가능성을 배제할 수 없으며, 보고된 일치율은 해당 개발·평가 공용 증빙 세트에 대한 결과로 해석한다. 아울러 기준값 평가자가 알고 있던 대상 시스템의 처리 전제가 프롬프트에 포함되지 않아 시스템의 판정이 기준값과 달라진 사례가 재실행에서 관찰되었는데(제6장 셋째 한계), 이는 판단 규칙이 기준값을 직접 재현하도록 구성되지 않았음을 보여준다. 이에 따른 해석 범위는 제6장 한계에서 다룬다.

오류의 유형과 환각을 판정하기 위해, 본 연구는 조작적 정의를 사전에 설정하였다. 판정 대상은 개별 사례(case)가 아니라 조서에 기재된 각 추출 필드 및 판단 근거 단위로 하였으며, 인간 평가자의 판정과 불일치하거나 조서 내 특정 사실·근거가 증빙과 대조되지 않는 항목을 대상으로 오류 유형을 네 가지로 분류하였다. 즉 ① 증빙에 존재하지 않는 정보를 사실처럼 생성한 경우를 ‘환각(hallucination)’, ② 정보 자체는 증빙과 일치하나 통제 기준의 적용에 실패한 경우를 ‘논리 연결 부족’, ③ 개별 사안의 맥락을 무시하고 일반 규칙을 기계적으로 적용한 경우를 ‘과도한 일반화’, ④ 증빙의 값을 잘못 판독한 경우를 ‘증빙 해석 오류’로 정의하였다. 이 분류 체계에서 환각은 ‘증빙에 근거가 없는 정보의 생성’으로 좁게 한정되며, 값의 오독(증빙 해석 오류)이나 기준 적용 실패(논리 연결 부족)와는 명확히 구분된다. 각 조서의 추출 필드와 판단 근거를 증빙과 대조하여 위 정의에 따라 환각 여부를 판정하였다.

2. Execution Efficiency and Accuracy

연구문제 (1)과 관련하여, 최초 1회 측정에서 시스템의 조서 초안 작성 시간은 809초로, 숙련 평가자의 21,300초 대비 96.2% 짧았다. 이는 증빙 수집·표본 추출과 산출물의 최종 검토·수정 시간을 제외한 측정 결과이며, 전체 감사 업무의 시간 절감률을 의미하지 않는다. 측정 구간은 생성형 AI의 경우 증빙 업로드 후 조서 산출 완료까지, 숙련 평가자의 경우 증빙 판독·입력 시작부터 조서 작성 완료까지이며, 두 경우 모두 증빙 수집과 표본 추출에 소요된 시간은 포함하지 않았다. 측정은 1회 수행하였으며, 동일 증빙에 대한 인간 평가자의 재측정은 학습 효과를 통제할 수

없이 반복 측정하지 않았다. 판정 정확성 측면에서, 인간 평가자의 판정을 기준값으로 한 오차행렬 분석 결과 전체 154건 중 152건이 일치하여 기준 평가자 판정과 일치율은 98.7%(95% 신뢰구간 [95.4%, 99.6%])였다(Table 2). 이 값은 PASS가 154건 중 148건(96.1%)을 차지하는 불균형한 정제 가상 데이터에서 단일 기준 평가자의 판정과 일치한 비율로, 일반적인 감사 정확도를 의미하지 않는다. PASS로 판정된 148건은 전건 정확히 식별되었으며, FAIL로 판정된 4건은 생성형 AI도 모두 FAIL로 판정하였다(재현율 4/4). 다만 결함 표본이 4건에 불과하여 확정적 성능이 아닌 초기 검증 결과로 해석되어야 한다(95% CI [39.8%, 100.0%]). 인간이 ‘확인필요’로 판정한 2건에 대해 생성형 AI는 ‘FAIL’로 판정하여 불일치가 발생하였다. 해당 2건은 모두 긴급 오류 수정으로 배포 후 승인이 이루어진 선배포·후승인 사례이며, 긴급 변경 맥락은 별도 속성에서 ‘확인필요’로 판정되어 있었다. 종합 판정을 FAIL로 만든 것은 테스트 수행 순서에 관한 속성으로, 해당 증빙에는 운영 테스트 부서가 공란인 반면 운영 테스터와 테스트일에는 값이 기재되어 있었고 그 시점이 요청 승인일보다 이른 상태였다. 세 평가자는 이 내부 불일치를 근거로 ‘확인필요’로 유보한 반면, 생성형 AI는 기재된 값에 판정 규칙을 기계적으로 적용하여 FAIL로 분류하였다(오류 유형 ③ 과도한 일반화). 기준값의 신뢰성을 확인하기 위해 3개 핵심 통제 60건에 대해 평가자 3인의 판정 일치도를 산출한 결과, 60건 전건에서 세 평가자의 판정이 일치하여 관찰 일치율은 100%, Fleiss's κ 는 1.000이었다. 특히 생성형 AI가 ‘FAIL’로 판정한 2건(REL-2025-001, REL-2025-006)에 대해 세 평가자 모두 독립적으로 ‘확인필요’로 판정하였으며, 세 평가자의 검토의견은 공통적으로 긴급 변경에 따른 선배포·후승인과 개발·운영 테스트 시점의 역전을 지적하였다. 이는 앞서 관찰된 불일치가 단일 평가자의 판단 성향에 기인한 것이 아니라, 증빙에 기재된 값에 판정 규칙을 기계적으로 적용한 시스템의 보수적 판정에서 비롯되었음을 뒷받침한다. 계정·권한 요청 승인 통제는 20건이 모두 PASS로 판정되어 단일 범주만 관찰됨에 따라 해당 통제의 κ 는 산출되지 않았으며, 위 값은 3개 통제를 통합한 60건 기준이다. 한편 본 실험에서 인간 평가자는 증빙에서 날짜·승인자·식별자 등의 정보를 직접 판독하여 입력한 뒤, 날짜 선후 관계 등의 정량적 대조에 한하여 엑셀 함수를 활용하는 반자동 방식으로 조서를 작성하였다. 증빙 판독과 입력 자체는 수작업으로 이루어졌으므로 제시된 인간 소요 시간은 순수 수작업 환경에 근접하며, 정량 대조를 자동화한 만큼 일반적인 전수 수작업보다 다소 보수적으로 측정된 값이다.

Table 2. Confusion Matrix (N=154)

	AI → PASS	AI → Uncertain	AI → FAIL
Human → PASS	148	0	0
Human → Uncertain	0	0	2
Human → FAIL	0	0	4

Note. Uncertain = "확인필요" (needs further review).

결함 식별 성능을 지표별로 살펴보면, 결함(FAIL) 탐지를 이진 문제로 볼 때 재현율(민감도)은 100%(4/4), 정밀도는 66.7%(4/6), 특이도는 98.7%(148/150), F1 점수는 80.0%로 산출되었다(Table 3). 다만 본 연구의 데이터는 PASS가 전체의 96.1%를 차지하는 극심한 클래스 불균형 구조를 가지므로, 전체 정확도 98.7%는 모든 사례를 PASS로 예측하는 단순 분류기의 정확도(96.1%)와 비교할 때 개선 폭이 2.6%p에 해당한다. 이러한 불균형의 영향을 보정하기 위해 다수 클래스에 가중되지 않는 지표를 함께 산출한 결과, 3개 판정(PASS/FAIL/확인필요)을 대상으로 한 Macro-F1은 0.600, Balanced Accuracy는 0.667이었다. 이는 인간이 '확인필요'로 판정한 2건을 AI가 모두 'FAIL'로 분류하여 해당 클래스의 재현율이 0%였던 데 기인하며, 결함 자체의 탐지(FAIL 재현율 100%)와 구분되는 한계이다. 또한 결함 표본이 4건으로 제한적이어서 추정 불확실성이 크다. 결함 재현율 100%의 정확검정(Clopper-Pearson) 95% 신뢰구간은 [39.8%, 100.0%]로 하한이 낮으므로, 재현율 수치는 확정적 성능이라기보다 결함 미탐지가 관찰되지 않았음을 보여주는 참고 지표로 해석해야 하며, 결함-경계 사례를 대폭 확대한 재검증이 필요하다. 감사 맥락에서 결함 누락(미탐지)이 결함 과잉 판정보다 훨씬 큰 위험임을 고려하면, 관찰된 보수적 판정 경향은 실무적으로 바람직한 방향일 수 있으나, 결함 재현율은 4건에 기반한 초기 검증 결과이므로 위 신뢰구간의 한계를 함께 고려하여 해석하여야 한다.

Table 3. Defect Detection Performance Metrics

Metric	Value	95% CI	Note
Accuracy	98.7%	[95.4%, 99.6%]	152/154, Wilson
Recall(FAIL)	100.0%	[39.8%, 100.0%]	4/4, Clopper-Pearson
Precision(FAIL)	66.7%	[22.3%, 95.7%]	4/6
Specificity	98.7%	[95.3%, 99.8%]	148/150
F1 Score(FAIL)	80.0%	-	-
Macro-F1	0.600	-	3-class
Balanced Accuracy	0.667	-	3-class

Note. FAIL detection treated as binary (FAIL vs. non-FAIL; Uncertain counted as non-FAIL); Macro-F1 and Balanced Accuracy over three classes (PASS/FAIL/Uncertain).

한편 단일 평가자 기준값의 타당성을 보완하기 위해, 위 3개 핵심 통제에서 총 60건을 원 평가자를 포함한 3인이 블라인드 조건에서 독립적으로 재평가하였다. 그 결과 세 평가자의 판정은 60건 전건에서 일치하였으며(단순 일치율 100%), 세 범주 전체에 대한 평가자 간 일치도는 Fleiss $\kappa=1.00$ 으로 완전 일치를 나타냈다(Table 4). 통제 별로는 프로그램 변경 승인과 계정·권한 회수에서 각각 $\kappa=1.00$ 이 산출되었고, 계정·권한 부여는 모든 판정이 PASS 단일 범주에 집중되어 우연 기대 일치도가 1에 수렴하므로 κ 를 산출하지 않고 단순 일치율로 보고하였다. 특히 판정이 쟁점이 되었던 사례들, 즉 프로그램 변경 통제의 '확인필요' 2건과 'FAIL' 2건, 계정·권한 회수 통제에서 퇴직 후 회수가 지연된 'FAIL' 1건에 대해 세 평가자의 판정이 모두 일치하여, 기준값이 복수 전문가 사이에서 재현 가능한 판정임을 확인하였다. 다만 세 평가자가 동일 법인·부서에 소속되어 동일 방법론을 공유하므로, 이는 조직 내 신뢰도 점검(within-organization reliability check) 수준의 근거로 해석되어야 한다.

Table 4. Inter-rater Agreement on Reference Judgments (3 raters, 60 cases)

Control	N	PASS	Uncertain	FAIL	Agreement	Fleiss κ
Program Change Approval	20	16	2	2	100%	1.00
Account/Privilege Grant	20	20	0	0	100%	– ¹
Account/Privilege Removal	20	19	0	1	100%	1.00
Total	60	55	2	3	100%	1.00

Note. ¹ κ not computed (all PASS, single category); reported as percent agreement. Overall κ computed over three categories across 60 cases.

개별 설계 요소가 판정에 미치는 영향을 탐색하기 위하여 절제 실험을 수행하였다. 동일 증빙과 동일 기반 모델, 동일 생성 파라미터(temperature=0) 아래에서 한 번에 한 축만 변경하는 조건을 구성하였으며, 변경 축은 프롬프트의 블록 구조, 통제별 판정 기준(케이스별 검토의견 문구 템플릿이 케이스 정의를 재노출하므로 함께 제거하였다), 증빙의 빈 셀 처리에 관한 제약, 권한 코드 정의표, 조직도, 출력 형식 지시, 자료형 안전 출력의 적용 여부, 그리고 모델 배정 방식이다. 교차 검증이 요구되는 3개 통제 146건을 대상으로 하였고, 판정 일치율에 관한 결과는 본 절에, 환각에 관한 결과는 다음 절에 제시한다. 다만 절제 실험은 최초 실험과 다른 시점의 단일 세션에서 수행되었으며, 다른 생성형 AI 방법이나 타 제공자 모델과의 기준선 비교는 포함하지 않았다. 따라서 조건 간 차이는 해당 세션 내의 상대 비교이며, 최초 실험의 98.7% 일치율이나 환각 미관찰이 개별 구성요소에 기인함을 보여주지 않는다.

절제 실험의 원 시스템 조건은 146건 중 122건이 일치하여 일치율이 83.6%였다. 이는 동일 사례에 대한 최초 실험의 98.6%(144/146)보다 15.1%p 낮았다. 판정 변동은 회수 이력 부재 시 케이스 분류를 확정할 수 없는 37건에 집중되었으나, 실행 시점에 따른 모델 변화와 프롬프트 기준의 공백이 각각 미친 영향은 분리하여 확인되지 않았다. 절제 실험 결과는 동일 세션 내 조건 간 비교로 해석한다.

Table 5. Comparison of Initial and Subsequent Evaluation Results

Evaluation	Scope	Conditions	Agreement with Reference Evaluator
Initial evaluation	154 cases across six controls	Original prompt	152/154 (98.7%)
Comparable subset of initial evaluation	146 cases across three controls	Original prompt	144/146 (98.6%)
Subsequent rerun	Same 146 cases across three controls	Stored original prompt	122/146 (83.6%)
Revised-prompt evaluation	50 cases (Account/Privilege Removal)	Explicit processing assumptions and rule priorities; three repeated runs	50/50 (100%) in each of three runs; judgments identical across runs ¹

Note. ¹ Supplementary observation on a single control; these results do not replace the figures reported in the text.

최초 실험, 후속 재실행 및 보완 프롬프트 실행의 평가 범위와 실행 조건, 기준 평가자와의 판정 일치 결과는 Table 5에 구분하여 제시하였다. 8-Block 구조를 제거하고 정보량을 동일하게 유지한 채 연속 산문으로 변환한 조건에서는 72.6%로 11.0%p 하락하였다. 동일한 146건에 두 조건을 적용한 대응 설계이므로 McNemar 정확검정을 적용한 결과, 8-Block 조건만 맞힌 사례가 17건, 산문 조건만 맞힌 사례가 1건으로 $p = .000145$ 였다. 통제별 판정 기준을 제거한 조건은 69.2%(14.4%p 하락), 권한 코드 정의표를 제거한 조건은 65.1%(18.5%p 하락), 조직도를 제거한 조건과 출력 형식 지시를 제거한 조건은 각각 73.3%(10.3%p 하락)였다.

자료형 안전 출력의 효과는 정확도가 아니라 산출 가능성에서 나타났다. 적용 조건은 146건 전건을 구조화된 조서 행으로 산출한 반면, 미적용 조건은 3개 통제 중 2개에서 행을 전혀 산출하지 못하였고 나머지 1개에서도 기준 50행 중 10행만 산출하였다. 구조와 스키마를 모두 제거한 최소 기준선 조건은 146건 중 30건만 산출하였으며, 산출된 30건에 한정하여 재채점한 경우에도 80.0%로 제안 방식에 미치지 못하였다.

모델 배정 방식과 관련하여, 경량 모델로 고정한 조건의 판정 일치율은 68.5%였으나 주목할 차이는 평균 정확도가 아니었다. 이 조건은 기준값의 결함 4건을 모두 탐지하지 못하여 결함 탐지 재현율이 0/4였던 반면, 고성능 모델 계열의 모든 조건은 4/4였다. 내역을 보면 4건 중 2건은 해

당 행 자체를 산출하지 못하였고 산출한 2건은 정상으로 판정하였으며, 전체적으로도 146건 중 40건에 대해 대응하는 판정을 산출하지 못하였다. 감사 맥락에서 결함 미탐지가 과잉 판정보다 위험한 오류 유형이므로, 모델 배경 정책은 평균 정확도가 아니라 결함 탐지 신뢰성과 출력 안정성을 기준으로 검토할 필요가 있다. 이 비교는 결함 4건에 근거한 관찰이므로, 배경 정책의 타당성은 결함 사례를 확대한 검증으로 확인되어야 한다.

동일 조건을 4회 반복 실행한 결과 146건 중 일치 건수는 각각 121건, 122건, 123건, 122건으로 변동 폭은 2건(1.3%p)이었다. 그러나 행 단위로는 회차 간 최대 15건의 판정이 달랐으며, 이는 두 판정이 모두 기준값과 어긋나 일치 건수에는 차이가 나타나지 않은 경우를 포함하기 때문이다. 이는 temperature=0 설정에서도 완전한 결정성이 보장되지 않는다는 앞선 서술을 뒷받침하는 실측치이다. 다만 이 반복은 동일 세션에서 원 시스템 조건만을 대상으로 한 4회 실행이므로, 출력 안정성을 일반적으로 보증하는 근거로 해석하지 않는다.

3. Hallucination Observation Results

연구문제 (2)와 관련하여, 스키마 검증과 코드 기반 검토의견 재작성을 거친 최종 조서 154건을 증빙과 대조한 결과, 정의된 기준에 따른 환각은 식별되지 않았다. 이는 후처리를 포함한 전체 파이프라인의 최종 산출물에 대한 관찰이며, 후처리 이전의 원시 AI 출력에서 환각이나 인용 오류가 발생하지 않았음을 의미하지 않는다. 다만 이는 해당 표본에서 정의된 기준에 따른 환각이 관찰되지 않았음을 의미할 뿐 발생 가능성이 0임을 보장하지는 않는다. 사건이 관찰되지 않은 표본에 대한 통계적 추정 기법인 rule of three[18]를 적용하면 95% 신뢰수준에서 발생률은 약 1.9%로 추정되나, 이는 각 사례가 독립적이라는 가정에 기반한 보조적 해석이다. 본 연구의 표본은 유사한 증빙 구조와 템플릿에서 파생되어 사례 간 완전한 독립성을 담보하기 어렵고 각 사례가 다수의 생성 필드를 포함하므로, 위 상한값은 참고치로 제한하여 해석한다.

아울러 환각 평가의 범위를 1단계 추출 구간까지 확장하였다. 테스트 내역 증빙 52건에 대해 판독 규칙 적용 여부를 달리한 두 조건으로 재추출하고, 증빙을 조인해 작성한 대조표와 필드 단위로 비교하였다. 364개 필드 중 360개가 일치하였으며 두 조건의 출력은 동일하였다. 불일치 4건은 모두 증빙 문서에 실재하는 값을 정확히 판독한 결과로, 해당 값이 대조표와 어긋난 것은 증빙 생성 단계의 오류에 기인한다. 따라서 추가로 검토한 테스트 내역 증빙

52건의 1단계 추출에서도 증빙에 근거가 없는 정보의 생성은 관찰되지 않았다. 다만 이 결과는 해당 증빙과 364개 평가 필드에 한정되며, 전체 통제의 모든 1단계 추출 결과에 대한 검증을 의미하지 않는다.

본 시스템은 단일 기제가 아니라 다음과 같은 여러 방어 장치를 결합한 다층 방어 구조를 채택하였으나, 환각 미관찰과 이 구조 사이의 인과관계는 본 실험에서는 확인되지 않는다. 첫째, 프롬프트 단계에서 확인 불가능한 필드는 임의 추정 없이 결측으로 처리하도록 제약하고, 증빙의 빈셀을 임의로 채우지 못하도록 행 번호와 식별자를 기준으로 셀 경계를 엄수시켰다. 둘째, 자료형 안전 출력 단계에서 결론값을 세 가지 열거형으로 제한하고 스키마를 위반한 출력을 즉시 폐기하여, 형식 오류나 임의 변형이 최종 산출물에 유입되는 것을 차단하였다. 셋째, 결정론적 문장 조합과 규칙 기반 후처리 검증을 통해 검토 의견이 자율 생성되며 오류가 유입되는 경로를 차단하였다. 다만 본 실험이 사전에 구조화·정제된 증빙을 사용하였으므로, 환각 미관찰의 원인을 방어 기제의 효과만으로 단정하기는 어려우며, 정제된 증빙 조건 자체가 환각 유발 요인을 제한했을 가능성도 함께 고려되어야 한다. 판독 규칙을 제거한 조건을 별도로 실행하여 방어 기제의 효과를 분리하고자 하였다. 그 결과 규칙 적용 여부에 따른 환각 발생의 차이는 관찰되지 않았다. 이는 방어 기제가 무효함을 의미한다기보다, 정제된 증빙 조건에서는 억제할 환각이 애초에 발생하지 않아 효과를 측정할 대상이 없었음을 시사한다. 이 효과의 검증은 비정제 증빙 환경을 대상으로 한 후속 연구에 남긴다.

4. Practitioner Survey(Acceptance Assessment)

연구문제 (3)과 관련하여, 실무 적용 가능성을 다각도로 평가하기 위해 ITGC 평가 실무 경험을 보유한 전문가를 대상으로 설문을 병행하였다. 응답자에게는 생성형 AI 솔루션의 시연과 산출 조서를 제시한 뒤 인식과 태도를 측정하였으며, 편의표집(convenience sampling)으로 총 30명의 응답을 확보하였다. 응답자 30명의 특성은 다음과 같다. IT 감사 및 ITGC 실무 경력은 3년 미만 13명, 3년 이상 7년 미만 10명, 7년 이상 12년 미만 4명, 12년 이상 3명으로 분포하였다. 소속은 대형 회계법인(Big 4 및 이에 준하는 규모)이 28명(93.3%)으로 대다수를 차지하였고, 나머지는 일반 기업의 내부감사 및 회계 부서였다. 주요 전문 분야는 외부감사(K-SOX 및 ITGC 독립적 검증)와 ITGC 구축·GRC(Governance, Risk, Compliance) 컨설팅이 중심을 이루었으며, 시스템 감리 및 보안 인증 심사

가 일부 포함되었다. 업무 목적의 생성형 AI 사용 빈도는 매일 사용이 18명으로 가장 많았고, 주 1~2회 7명, 월 1~2회 3명, 시도 경험 1명, 미사용 1명이었다. 연령대는 20대 10명, 30대 18명, 40대 1명, 50대 이상 1명이었다. 다만 편의표집에 따른 표본의 대표성 한계와 상대적으로 작은 표본 규모를 고려하여, 설문 결과는 확정적 일반화보다 탐색적 해석에 초점을 두었다. 본 설문은 개인식별정보를 수집하지 않는 익명 조사로 수행하였으며, 응답 전에 연구 목적과 자료 처리 방침을 설명하고 참여자의 자발적 동의를 얻었다. 기관생명윤리위원회의 사전 심의는 받지 않았다. 다만 향후 연구에서는 사전 윤리 심의 절차를 명시적으로 밝을 필요가 있다. 측정 도구는 선행연구의 척도를 참고하여 각 구성개념을 7점 리커트 척도의 복수 문항으로 측정하였으며, 핵심 구성개념은 산출물 신뢰성 5문항, 실무 활용성 5문항, 사용 의도 4문항, 전문적 의구심 6문항, HITL 개입 적정성 5문항으로 측정하였다. 선행연구 척도를 참고한 구성개념의 출처는 [19]-[22]와 같다. 역문항은 역채점한 후 구성개념 점수를 산출하였다. 산출물 신뢰성은 5문항 중 4문항 이상, 사용 의도는 4문항 중 3문항 이상에 유효하게 응답한 경우 유효 응답 문항의 평균으로 산출하였다. ‘모름’ 및 무응답은 결측으로 처리하여 평균 산출에서 제외하였다. 각 구성개념의 기술통계는 Table 6과 같다. 최종 분석 문항을 기준으로 산출한 Cronbach's α 는 산출물 신뢰성 .939, 실무 활용성 .910, 사용 의도 .963, 전문적 의구심 .940이었다. 신뢰도 분석은 각 척도의 모든 문항에 유효하게 응답한 사례만 포함하였으며, 산출물 신뢰성과 사용 의도는 각각 29명, 실무 활용성·전문적 의구심·HITL 개입 적정성은 각각 30명이었다. 반면 산출물 신뢰성과 사용 의도의 척도 평균 산출 기준을 충족한 응답자는 각각 30명으로, 기술통계에는 모두 포함하였다. HITL 개입 적정성은 연구자가 개발하고 정보시스템 감사 통제 분야 연구자의 검토를 거친 5개 문항으로 측정하였으며, Cronbach's α 는 .814였다. 척도의 구성타당도는 별도로 검증하지 않았으므로 탐색적 지표로 해석한다. 핵심 5개 차원의 기술통계 결과(Table 6)에서 전문적 의구심(M=5.82)이 가장 높게 나타난 반면 산출물 신뢰성(M=4.14)이 가장 낮았다.

Table 6. Descriptive Statistics of AI Perception and Attitude

Type	Dimension	N	M	SD
Core	Skepticism	30	5.82	1.41
Core	HITL Intervention	30	5.42	1.25
Core	Practical Usefulness	30	5.22	1.45
Core	Intention to Use	30	5.19	1.67
Core	Output Trust	30	4.14	1.41
Aux.	Manual Limitation	30	5.29	1.26
Aux.	AI Replacement	29	4.95	1.50
Aux.	Control Fitness	30	4.79	1.41
Aux.	Human-AI Comparison	30	3.97	1.21

Note. Core = core dimensions addressing RQ3; Aux. = auxiliary dimensions for exploratory analysis; N = number of respondents with valid scale or item scores under the scoring criteria described in the text; M = mean; SD = standard deviation. One respondent did not answer the AI Replacement item (N=29).

아울러 생성형 AI가 작성한 조서의 품질을 통제별로 평가하기 위해, 6개 통제 각각에 대해 정확성·완전성·명확성·논리성·실무 적합성의 5개 항목을 7점 척도로 측정하였다. 조서 품질 종합 평균은 5.56(SD=1.28)으로 양호한 수준이었으며, 단순 파라미터 대조 통제인 비밀번호 설정(M=5.91)과 환경 분리(M=5.87)가 가장 높고, 교차 검증 복잡도가 높은 프로그램 변경(M=5.30)과 계정·권한 회수(M=5.39)가 상대적으로 낮게 평가되었다(Table 7).

Table 7. AI Workpaper Quality by Control (7-point scale, N=30)

Control	M	SD
Password Setting	5.91	1.39
Env. Segregation	5.87	1.45
Access Request Approval	5.46	1.45
Privileged Access	5.43	1.40
Access Removal	5.39	1.31
Program Change Approval	5.30	1.37
Overall	5.56	1.28

대응표본 t-검정 결과 산출물 신뢰성과 실무 활용성(M=5.22) 간에 유의한 차이가 확인되었으며($t_{29}=-4.822$, $p<.001$, Cohen's $d=-0.88$)[23], 전문적 의구심(M=5.82)과 HITL 개입(M=5.42) 간에도 유의한 평균 차이가 나타나 의구심 유지 요구가 인간 개입 필요성 인식보다 높게 나타났다($t_{29}=2.759$, $p=.010$, Cohen's $d=0.50$). 주요 차원 간 검정 결과는 Table 8과 같다. 다만 산출물 신뢰성, 실무 활용성, 전문적 의구심, HITL 개입은 서로 다른 구성개념

이므로, 이들 간의 평균 비교는 척도 점수의 차이를 기술적으로 제시하기 위한 것이며 구성개념 간 대소를 규범적으로 단정하기 위한 것은 아니다. 본 비교는 산출물 신뢰성의 평가 점수가 실무 활용성보다 상대적으로 낮은 응답 패턴을 탐색적으로 보여주는 보조 분석이다. 신뢰성 평균 4.14는 7점 척도의 중간값 4보다 약간 높으므로, 이를 절대적으로 낮은 신뢰로 해석하지 않는다.

Table 8. Paired-Samples t-test Results

Comparison	Trust vs. Use	Intent vs. Trust	Skepticism vs. HITL
M1	4.14	5.19	5.82
M2	5.22	4.14	5.42
Diff	-1.08	+1.05	+0.40
t	-4.822	3.943	2.759
df	29	29	29
p	<.001	<.001	.010
Cohen's d	-0.88	0.72	0.50

Note. M(1) and M(2) = means of the first and second dimensions in each comparison; Diff = mean difference; d = Cohen's d. Each paired-samples t-test included 30 respondents with valid scores on both scales being compared (df = 29). Scale scores were calculated according to the scoring criteria described in the text.

이는 실무자들이 생성형 AI의 유용성은 인정하면서도 산출물 자체의 신뢰에는 유보적인 양면적 태도를 보임을 보여준다. 아울러 Pearson 상관분석 결과 전문적 의구심과 HITL 개입 필요성 인식 간에는 강한 정적 상관관계가 관찰되었다($r=.827$, $p<.01$). 이는 두 인식이 함께 높게 나타나는 경향을 보여주며, 횡단면 설문 결과만으로 인과관계나 영향 방향을 확인할 수는 없다. 다만 표본 규모($N=30$)를 고려하여 상관분석 결과는 관계의 방향성과 강도를 파악하는 탐색적 수준에서 해석하였다.

이러한 결과는 실무자들이 생성형 AI를 유용한 도구로 활용하면서도 그 산출물을 무조건적으로 신뢰하지는 않는 양면적 태도로 수용하고 있음을 보여준다. 이러한 '낮은 신뢰-높은 활용' 패턴의 이론적 해석과 실무적 함의는 제5장에서 논의한다.

한편 연령 집단 간 비교에서도 유의한 차이가 관찰되었다. 표본 수가 적은 40대 이상을 30대 이상 그룹으로 합산하여 20대와 비교한 결과, 전문적 의구심에서 20대($n=10$, $M=6.72$)가 30대 이상($n=20$, $M=5.38$)보다 유의하게 높았으며(Mann-Whitney $U=174.5$, $p<.001$, $r=.60$), 인간 개입 필요성 인식에서도 20대($M=6.14$)가 30대 이상($M=5.06$)보다 유의하게 높았다($U=163.0$, $p=.004$, $r=.51$). 효과크기 r 은 $Z/\sqrt{N}$ 으로 산출하였다. 이는 20대 집단이

생성형 AI 산출물에 대해 더 강한 비판적 시각을 유지하고 인간 개입의 필요성을 더 높게 인식하는 경향을 시사한다. 다만 연령 집단 비교는 다중 비교에 대한 보정 없이 수행하였고 표본 규모($N=30$)가 작으므로 탐색적 결과로 해석하며, 연령은 실무 경력과 같은 변수가 아니므로 이를 경력 차이로 해석하지 않는다.

V. Discussion

본 연구는 정제된 가상 증빙에 대한 최초 단회 실험에서 시스템이 조서 초안 작성 시간을 단축하고 숙련 평가자의 판정과 높은 일치율을 보였음을 확인하였다. 이 일치율은 PASS가 대부분인 불균형 데이터에서 단일 기준 평가자와 일치한 비율이며, 결함 탐지 결과는 FAIL 4건에 근거한 초기 검증 결과이다. 다만 재실행 성능 차이와 개발·평가 자료의 미분리를 고려하면, 이를 숙련 평가자와 동등한 정확성이나 실제 감사 환경의 성능으로 일반화하기는 어렵다. 구조화된 프롬프트와 자료형 안전 출력 및 후처리를 적용한 최종 산출물에서는 정의된 환각이 관찰되지 않았다. 다만 비교 조건에서도 환각이 관찰되지 않아, 개별 설계 요소의 환각 억제 효과는 확인하지 못하였다. 따라서 본 결과는 해당 실험 조건에서의 관찰로 한정하며, 비정제 증빙 환경에서 추가 검증이 필요하다.

한편 숙련 평가자가 예외 식별에서 우위를 보인 사례는, 생성형 AI의 근본적 판단 한계라기보다 맥락 정보가 판정 단계에 연계되는 방식에 개선 여지가 있음을 보여준다. 불일치한 2건에서 생성형 AI는 명시적 증빙의 정량적 대조(날짜·승인자 등)는 정확히 수행하였으나, 증빙 내부의 값 불일치(운영 테스트 부서는 공란이나 테스터와 테스트일은 기재된 상태)를 판단 유보의 사유로 다루지 못하고 기재된 값에 판정 규칙을 기계적으로 적용하였다. 이는 입력 스키마와 판정 규칙의 개선을 통해 보완될 수 있는 부분이다. 다만 이러한 불일치가 결함 누락이 아니라 보수적 과잉 판정의 형태로 나타났다는 점은, 감사 관점에서 상대적으로 안전한 오류 유형에 해당한다. 확인필요 사례를 FAIL로 분류하는 보수적 판정은 결함 누락을 줄이는 방향으로 작용할 가능성이 있으나, 추가 검토 부담과 불필요한 결함 판단을 유발할 수 있다. 따라서 그 실무적 적절성은 미탐지 위험과 과잉 경고 비용을 함께 평가하여 판단해야 한다.

실무자 설문에서는 실무 활용성에 비해 산출물 신뢰성의 평가 점수가 상대적으로 낮게 나타났다. 이러한 응답 패턴은 전문 영역의 AI 수용을 기술 성능과 자동화 신뢰

관점에서 함께 검토할 필요성을 시사한다[15]. 제조 중소기업의 AI 솔루션 구축의도를 분석한 연구에서도 정보품질, 시스템품질 및 조직환경과 구축의도의 관계를 인지된 유용성과 신뢰성을 포함한 모형으로 검토하였다[24]. 실무자들은 생성형 AI의 유용성을 인정하면서도 산출물을 무조건 신뢰하기보다 인간 검토를 전제로 한 조건부 수용의 태도를 보였으며, 전문적 의구심이 높을수록 인간 개입의 필요성을 크게 인식하였다. 이 결과는 전문 영역에서 생성형 AI의 활용이 기술 성능만으로 결정되지 않으며, 신뢰성 확보를 위한 검증 절차와 인간의 감독 체계가 함께 구축될 때 실무 적용 가능성이 높아질 수 있음을 시사한다. 즉 생성형 AI의 성능 향상뿐 아니라 인간의 감독 체계를 함께 설계하는 것이 전문 영역의 AI 도입을 위한 핵심 조건이다. 본 연구의 결과는 HITL이 구체적으로 어느 단계에 배치되어야 하는지에 대한 실무적 함의를 제공한다. 제안 시스템은 판정 결과를 PASS, FAIL, 확인필요의 세 가지로 산출하는데, 인간 검토는 이 판정 유형에 따라 차등적으로 적용될 수 있다. 첫째, 모델이 판단 근거가 부족하다고 판단한 ‘확인필요’ 건은 자동 승인 없이 반드시 인간 평가자의 검토로 이관된다. 둘째, 실험에서 관찰된 바와 같이 생성형 AI가 인간의 ‘확인필요’ 판정을 ‘FAIL’로 과잉 판정하는 보수적 오류가 발생할 수 있으므로, FAIL로 판정된 건에 대해서도 인간이 맥락을 고려하여 최종 확정하는 절차가 필요하다. 셋째, PASS로 판정된 건의 표본 기반 사후 점검은 가능한 운영 방안으로 제안한다. 다만 결함 표본이 제한적이고 실제 환경의 미탐지 위험을 검증하지 않았으므로, 표본 점검의 충분성 및 검토 부담 감소 효과는 별도의 실증 검증이 필요하다. 판정 신뢰도가 낮거나 오류 위험이 큰 지점에 인간 검토를 선별적으로 배치하는 협업 구조를 제안한다. 따라서 실무 도입을 평가할 때는 시스템 성능과 함께 인간 검토의 배치, 책임 범위 및 검토 부담을 고려할 필요가 있다. 위와 같은 판정 유형별 차등 검토 방안은 본 연구가 제안하는 설계 수준의 협업 구조로서, 실제 오류 감소나 검토 효율에 미치는 효과는 본 연구에서 실험적으로 검증되지 않았으며 향후 연구를 통해 확인되어야 한다. 한편 본 시스템을 실제 감사 환경에 도입하기 위해서는 보안과 기밀성에 대한 고려가 선행되어야 한다. 실제 ITGC 증빙에는 계정 정보, 조직도, 권한 내역, 변경 이력 등 민감정보가 포함되므로, 외부 생성형 AI API로 증빙을 전송하는 구조는 데이터 보호 관점에서 신중한 검토가 필요하다. 구체적으로 전송 구간의 암호화, 입력 데이터의 모델 학습 사용 배제, 데이터 보존 기간과 저장 위치(데이터 지역성)의 통제, 접근 권한 관리, 호출 및 처리 이

력에 대한 로그와 감사 추적의 확보 등이 요구된다. 본 연구는 통제된 실험 환경에서 가상 증빙을 사용하였으므로 이러한 보안 요건을 실증적으로 다루지 않았으나, 실무 적용 단계에서는 온프레미스 또는 프라이빗 배포, 민감정보 비식별화, 계약 수준의 데이터 처리 보증 등을 통해 위 요건을 충족하는 설계가 병행되어야 한다.

기업별 보안 정책과 기밀 유지 의무로 인해 실제 ITGC 증빙을 연구 목적으로 확보하거나 외부 생성형 AI API에 전송하는 데에는 제약이 있을 수 있다. 이러한 제약을 고려할 때, 본 연구는 기업 기밀정보의 제공 없이 동일한 가상 증빙에 대한 인간-AI 비교, 반복 실험 및 절제 실험을 수행하여 시스템의 성능과 취약점을 점검했다는 강점을 가진다. 이는 현장 도입에 앞서 활용할 수 있는 사전 검증틀로 의의가 있으며, 실제 감사 환경에서의 성능과 보안 적합성은 후속 검증이 필요하다.

VI. Conclusions

본 연구는 ICFR 내 ITGC 평가조서 작성을 자동화하는 생성형 AI 모델을 제안하고, 통제된 가상 증빙 조건에서 숙련된 평가자의 산출물과 비교하여 실무 적용 가능성을 탐색한 초기 실증 연구이다. 최초 단회 실험에서 시스템은 조서 초안 작성 시간을 96.2% 단축하고, 154건에서 단일 숙련 평가자의 판정과 98.7% 일치하였다. 이는 154건 중 148건이 PASS인 불균형한 정제 가상 데이터에서의 일치율로 일반적인 감사 정확도를 의미하지 않으며, FAIL 4건을 모두 탐지한 결과도 확정적 성능이 아닌 초기 검증 결과이다. 후처리 완료 산출물에서는 정의된 기준에 따른 환각이 식별되지 않았다. 다만 동일한 3개 통제 146건의 후속 재실행 일치율은 83.6%로 나타났다. 최초 결과는 실험 조건과 판정 기준에 의존하는 관찰로 해석한다. 반면 숙련 평가자는 암묵적 맥락과 조직 관행이 요구되는 예외 식별에서 우위를 보였고, 실무자 설문에서는 산출물 신뢰성의 평가 점수가 실무 활용성보다 유의하게 낮게 나타났다. 이는 활용성에 비해 상대적으로 낮은 신뢰를 보여주는 탐색적 응답 패턴으로 해석한다.

본 연구의 기여는 다음 세 가지로 요약된다. ① 이론적으로, 본 연구는 ITGC 평가라는 구체적 실무 단위에서 생성형 AI와 숙련 평가자를 동일 조건에서 직접 비교한 실증 근거를 제시하였다. ② 방법론적으로, 8-Block 구조화 프롬프트와 자료형 안전 출력을 결합한 다층 방어 설계를 제안하고, 통제된 조건에서 환각 발생을 정량적으로 관찰-검

증하는 절차를 제시하였다. 후속 재실행 세션의 절제 비교에서는 8-Block 구조화 프롬프트(-11.0%p, McNemar 정확검정 $p=.000145$), 통제별 판정 규칙(-14.4%p), 권한 코드 정의표(-18.5%p)를 각각 제거한 조건의 일치율이 하락하였고, 자료형 안전 출력을 적용하지 않은 조건에서는 조서 행이 대부분 산출되지 않았다. 다만 이는 원 시스템 조건의 일치율이 83.6%였던 단일 세션 내 상대 비교이며 타 방법·타 제공자 모델과의 기준선 비교를 포함하지 않으므로, 최초 실험의 높은 일치도가 개별 구성요소에 기인한다고 해석하지 않는다. 반면 환각 억제 효과는 어느 조건에서도 조건 간 차이로 관찰되지 않았다. 이는 비교한 조건의 평가 산출물에서 정의된 환각이 관찰되지 않아 조건 간 억제 효과를 추정할 수 없었기 때문이며, 설계 요소가 환각 억제에 미치는 인과적 효과의 규명은 비정제 증빙 환경에서의 후속 검증 과제로 남는다. ③ 실무적으로, 성능 지표와 실무자 수용성을 함께 분석함으로써 신뢰성 있는 AI 도입을 위한 검증 프레임워크 설계의 기초 자료를 제공한다. 또한 기업 기밀정보로 인해 실제 감사 증빙의 확보와 외부 전송이 제한될 수 있는 상황에서, 가상 증빙을 활용한 인간-AI 비교와 반복 실행 및 절제 실험을 통해 현장 도입 전 시스템의 성능과 취약점을 점검하는 사전 검증 틀을 제시하였다.

특히 본 연구는 생성형 AI의 성능 검증에만 초점을 두지 않고, 실무자의 수용성과 인간 개입(HITL)의 필요성을 함께 분석함으로써 IT 감사 자동화 시스템의 설계와 검증 절차에 활용할 수 있는 실무적 근거를 제시하였다. 또한 생성형 AI의 성능과 실무 수용성을 하나의 연구 틀에서 함께 탐색한 초기 실증연구로서, 관련 분야의 후속 연구를 위한 실증적 기반을 제공하고 생성형 AI 기반 IT 감사 자동화의 실무 확산에 기여할 수 있는 토대를 마련하였다.

본 연구는 다음과 같은 한계를 가진다. 첫째, 판정의 기준값을 1인의 숙련된 평가자가 작성한 조서로 설정하였으므로, 결과의 정확도는 해당 기준의 타당성을 전제로 해석되어야 한다. 3개 핵심 통제 60건에 대한 3인 평가자 일치도 검토로 기준값의 신뢰성을 부분적으로 확인하였으나, 재평가 표본의 91.7%가 PASS로 판정된 불균형 구조이므로 경계 사례를 확대한 검증과 조직 외부 평가자를 포함한 교차 검증은 후속 과제로 남는다. 둘째, 본 결과는 사전에 구조화·정제된 증빙에 기반한 조건부 관찰로서 비정제 증빙이 혼재된 실제 감사 환경으로의 일반화에는 제약이 있다. 셋째, 실험은 단일 모델 계열과 특정 파이프라인 조건에서 수행되었으므로 생성형 AI 전반으로의 일반화에는 추가 검증이 필요하다. 또한 앞서 기술한 바와 같이 모델

수준의 완전한 재현성은 보장되지 않는다. 저장된 프롬프트를 그대로 사용하여 반복 실행한 결과, 판정의 변동은 전체에 고르게 나타나지 않고 특정 조건에 집중되었다. 3개 통제 146건 가운데 124건은 반복 실행과 4개월 후 재실행에서 모두 동일한 판정이 산출되었으며, 판정이 달라진 22건 중 21건은 부서이동 발령자 가운데 권한 회수 이력에 대응 레코드가 없어 케이스 분류 기준의 성립 요건을 확정할 수 없는 37건 구간에 속하였고 나머지 1건은 프로그램 변경 통제의 사례였다. 해당 구간은 회수 대상 권한이 없는 발령자에게 회수 이력이 생성되지 않는다는 대상 시스템의 처리 전제가 프롬프트에 포함되지 않아, 판정 기준이 결론을 확정하지 못하는 상태였다. 처리 전제와 판정 규칙의 우선순위를 명시한 보완 조건으로 계정·권한 회수 통제 50건을 3회 실행한 결과, 세 회차의 판정은 행 단위로 동일하였고 모두 50건이 기준값과 일치하였다. 다만 이는 단일 통제에 대한 추가 관찰이며, 본문에 보고한 수치는 보완 이전의 프롬프트로 측정한 값이다. 본 실험 기록에서는 판정 변동이 케이스 분류 기준에 공백이 있는 구간에 집중되었다. 이는 규칙의 명확성이 출력 안정성과 관련될 가능성을 시사하지만, 모델 변화와 실행 조건의 영향을 분리하지 못했으므로 변동 원인을 단정하기는 어렵다. 재현성을 높이기 위해서는 케이스 정의와 판정 규칙의 우선순위를 명확히 하고, 가능한 경우 모델 스냅샷을 고정하여 기록해야 한다. 입력 자료, 프롬프트 버전, 호출 설정 및 후처리 코드의 버전도 함께 관리할 필요가 있다. 넷째, 본 연구는 한국의 K-SOX 체계를 배경으로 한국어 증빙과 국내 실무자를 대상으로 수행되었으므로, 단일 국가·단일 언어 환경이라는 제약을 가지며 다른 규제 체계나 언어 환경으로의 일반화에는 추가 검증이 요구된다. 다섯째, 설문 표본($N=30$)의 규모를 고려할 때 설문 분석 결과는 탐색적 수준에서 해석되어야 한다. 따라서 서로 다른 상용·공개 대규모 언어모델 간의 성능 비교, 비정제 증빙 환경에서의 다층 방어망 검증, 결함·경계 사례를 확대한 재검증, 다국어·다국가 환경으로의 확장, 그리고 대규모 표본을 통한 수용성 분석은 후속 연구의 과제로 남는다. 여섯째, 기준값에서 FAIL로 판정된 결함 사례는 4건이고, 별도로 확인 필요 사례가 2건이었다. 결함 및 판단 유보 사례가 모두 제한적이므로, 결함 탐지 성능과 불확실성 분류 성능의 통계적 안정성에 제약이 있다. 일곱째, 개발용 증빙과 평가용 증빙을 분리하지 않았고 판단 규칙과 증빙의 설계자가 기준값 평가자와 동일인이다. 따라서 보고한 98.7%는 미지의 증빙에 대한 일반화 성능이 아니라 판단 규칙이 해당 증빙 세트에 적합한 상태에서의 일치율이며, 분리된 세트

에 의한 검증은 후속 과제로 남는다. 아울러 프롬프트에 통제별 판단 논리가 포함되어 있어 평가 과제가 실제 감사보다 단순화되었을 가능성이 있다. 통제별 판정 규칙을 제거한 조건을 실행한 결과, 판정 일치율이 83.6%에서 69.2%로 14.4%p 하락하였다. 그러나 하락의 양상은 결함 탐지의 실패가 아니라 정상 건에 대한 과잉 경보였다. 결함 4건에 대한 재현율은 두 조건 모두 100%였던 반면, FAIL 판정의 정밀도는 57.1%에서 8.3%로, 특이도는 97.9%에서 69.0%로 하락하였다(FAIL로 판정한 건수 7건 → 48건, 그중 실제 결함은 두 조건 모두 4건). 이는 통제별 판단 논리가 결함의 소재를 지시하기보다 정상으로 처리할 범위의 경계를 규정하는 기능을 수행하였음을 시사한다. 다만 결함 표본이 4건에 불과하여 재현율 비교의 검정력은 제한적이다. 여덟째, 절제 실험은 후속 재실행 시점의 단일 세션에서 원 시스템 조건 대비 상대 비교로 수행되었고, 다른 생성형 AI 방법이나 타 제공자 모델과의 기준선 비교는 수행하지 않았다. 비교 대상 모델도 단일 제공자의 두 모델에 한정되며, 반복 실행 역시 원 시스템 조건 4회에 그쳤다. 따라서 프롬프트 구조·자료형 안전 출력·모델 라우팅의 개별 효과는 제한적으로만 관찰되었으며, 최초 실험의 일치도나 환각 미관찰에 대한 인과적 기여로 해석할 수 없다. 아홉째, 스키마 검증을 통과한 최종 산출물만을 평가 대상으로 하였다. 본 시스템은 스키마 위반 시 재생성을 시도하지 않으므로 재생성에 따른 선택 효과는 존재하지 않는다. 절제 실험에서 계측한 전체 호출 통계는 총 117회, 최초 성공 100회, 스키마 검증 실패 15회, 재생성 0회, 최종 실패 2회(경량 모델 조건)였다. 또한 조서 품질은 실무자 설문으로 평가하였으며, 개별 필드의 정확도를 독립 전문가가 증빙과 대조하여 평가하는 블라인드 평가는 수행하지 않았다(필드 단위 대조는 테스트 내역 증빙 52건의 1단계 추출에 한정). 열째, 실제 감사 현장에서 사용자가 본 시스템을 활용하여 업무를 수행하는 상황에 대한 검증은 이루어지지 않았다. 열한째, 본 연구는 가상 증빙을 사용하여 외부 API 전송에 따른 보안·기밀성 요건(암호화, 데이터 보존·지역성, 학습 사용 배제, 접근통제, 감사 추적 등)을 실증적으로 검증하지 않았으므로, 민감정보가 포함된 실제 환경으로의 확장을 위해서는 별도의 보안 설계와 검증이 요구된다.

REFERENCES

- [1] Gartner, "Forecast: AI Spending, Worldwide, 2025-2030, 1Q26," Gartner, Inc., 2026.
- [2] F. Dell'Acqua, E. McFowland III, E. R. Mollick, H. Lifshitz-Assaf, K. Kellogg, S. Rajendran, L. Kraye, F. Candelon, and K. R. Lakhani, "Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality," SSRN Electronic Journal, 2023. <https://doi.org/10.2139/ssrn.4573321>
- [3] J. Kokina and T. H. Davenport, "The emergence of artificial intelligence: How automation is changing auditing," *Journal of Emerging Technologies in Accounting*, Vol. 14, No. 1, pp. 115-122, 2017. <https://doi.org/10.2308/jeta-51730>
- [4] M. Eulerich, A. Sanatizadeh, H. Vakilzadeh, and D. A. Wood, "Is it all hype? ChatGPT's performance and disruptive potential in the accounting and auditing professions," *Review of Accounting Studies*, 2024. <https://doi.org/10.1007/s11142-024-09833-9>
- [5] S. Emett, M. Eulerich, E. Lipinski, N. Prien, and D. A. Wood, "Leveraging ChatGPT for enhancing the internal audit process—A real-world example from Uniper," *Accounting Horizons*, Vol. 39, No. 2, pp. 125-135, 2025. <https://doi.org/10.2308/HORIZONS-2023-111>
- [6] M. M. Dong, T. C. Stratopoulos, and V. X. Wang, "A scoping review of ChatGPT research in accounting and finance," *International Journal of Accounting Information Systems*, Vol. 55, 100715, 2024. <https://doi.org/10.1016/j.accinf.2024.100715>
- [7] Samil PwC, "Samil PwC develops 'K-SOX Hub,' an AI-based solution for internal control over financial reporting," Samil PwC Press Release, 2025. (in Korean)
- [8] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, Vol. 55, No. 12, pp. 1-38, 2023. <https://doi.org/10.1145/3571730>
- [9] ISACA, COBIT 2019 Framework: Governance and Management Objectives, Rolling Meadows, IL: ISACA, 2018.
- [10] PCAOB, AS 2201: An Audit of Internal Control Over Financial Reporting That Is Integrated with an Audit of Financial Statements, Public Company Accounting Oversight Board, 2007.
- [11] KPMG, "2023 SOX Report," KPMG LLP, 2023.
- [12] A. Fedyk, J. Hodson, N. Khimich, and T. Fedyk, "Is artificial intelligence improving the audit process?," *Review of Accounting Studies*, Vol. 27, No. 3, pp. 938-985, 2022. <https://doi.org/10.1007/s11142-022-09697-x>
- [13] M. Eulerich, A. Masli, J. Pickerd, and D. A. Wood, "The impact of audit technology on audit outcomes," *Contemporary Accounting Research*, Vol. 40, No. 2, pp. 981-1012, 2023. <https://doi.org/10.1111/1911-3846.12847>
- [14] B. E. Christensen, S. A. Emett, M. Eulerich, and D. A. Wood,

- "Using artificial intelligence for fraud risk assessment: Evidence from novice and experienced auditors," SSRN Electronic Journal, 2025. <https://doi.org/10.2139/ssrn.5173407>
- [15] R. Parasuraman and D. H. Manzey, "Complacency and bias in human use of automation: An attentional integration," *Human Factors*, Vol. 52, No. 3, pp. 381-410, 2010. <https://doi.org/10.1177/0018720810376055>
- [16] S. Schulhoff, M. Ilie, N. Balepur, K. Kahadze, A. Liu, C. Si, et al., "The prompt report: A systematic survey of prompting techniques," *arXiv*, 2024. <https://doi.org/10.48550/arXiv.2406.06608>
- [17] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha, "A systematic survey of prompt engineering in large language models: Techniques and applications," *arXiv*, 2024. <https://doi.org/10.48550/arXiv.2402.07927>
- [18] J. A. Hanley and A. Lippman-Hand, "If nothing goes wrong, is everything all right? Interpreting zero numerators," *JAMA*, Vol. 249, No. 13, pp. 1743-1745, 1983. <https://doi.org/10.1001/jama.1983.03330370053031>
- [19] R. K. Hurtt, "Development of a scale to measure professional skepticism," *Auditing: A Journal of Practice & Theory*, Vol. 29, No. 1, pp. 149-171, 2010. <https://doi.org/10.2308/aud.2010.29.1.149>
- [20] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly*, Vol. 13, No. 3, pp. 319-340, 1989. <https://doi.org/10.2307/249008>
- [21] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MIS Quarterly*, Vol. 27, No. 3, pp. 425-478, 2003. <https://doi.org/10.2307/30036540>
- [22] M. Madsen and S. Gregor, "Measuring human-computer trust," in *Proc. of the 11th Australasian Conference on Information Systems*, Brisbane, Australia, pp. 6-8, 2000.
- [23] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed., Lawrence Erlbaum Associates, 1988.
- [24] D. Lee, Y. Park, G. Oh, and H. Lee, "Comparative Analysis of AI Adoption Strategies Between Manufacturing Firms and Solution Providers," *Journal of The Korea Society of Computer and Information*, Vol. 30, No. 10, pp. 231-242, 2025. <https://doi.org/10.9708/jksci.2025.30.10.231>

Authors



Jisu Sohn received the M.S. degree in Digital Management from Korea University, Korea, in 2026. She has professional experience in security vulnerability assessment, penetration testing, and IT auditing, with a focus on IT

General Controls (ITGC) evaluation. Her research interests include generative AI, IT audit automation, and internal control over financial reporting.



Kyuhwan Lee is currently a software developer in the banking industry, where he designs and implements enterprise software systems for core banking and financial services. He previously worked as an IT

security consultant, focusing on vulnerability assessment and security compliance. His research interests include artificial intelligence, information security, and financial software systems.



YoungGeun Park received a Ph.D. degrees in Digital Management, Korea University in 2025. He holds a bachelor's degree in International Trade and Logistics, a master's degree in Marketing from Korea University's Graduate

School of Business. His main areas of interest include online food sales strategies, distribution innovation, and digital marketing.



Dongsu Lee received the Ph.D. degree in Digital Management from Korea University in 2025. He is currently a Professor and Director of the Network Center at the School of Medicine and Science, Gachon University.

He previously held executive positions at major IT companies, including LG CNS, TmaxSoft, and Daewoo Information. His research interests include artificial intelligence in manufacturing and home healthcare.

비교	공통	불일치	비교	공통	불일치
C4 ↔ C4+3	146	0	C4 ↔ C9	146	15 (전부 IT01.R02 계정-관련 회수)
C4 ↔ C4+2	146	1 (IT02.R05 프로그램 변경)	C4+3 ↔ C9	146	15 (전부 IT01.R02 계정-관련 회수)

2026년 5월 운영 실행과 2026년 9월 재실행의 비교도 같은 양상이다.

종재	공통	불일치
IT02.R05 프로그램 변경	52	1
IT01.R01 계정-관련 부여	44	0
IT01.R02 계정-관련 회수	50	21

146건 중 109건은 모든 실행에서 동일한 편향이 산출되었다. 혼동된 행은 예외 없이 계정-관련 회수 총재의 발령구분=부서이동*이면서 권한행=미확인인 37건 구간 내에 있다.

F-2. 그 구간에서 편향 기준이 결론을 확정하지 못하는 이유

[고지함은 2단계 제이스 분류 및 A/B 판정 기준]
권한행은 [계정/관련 회수 이력]에서 가져오며, 발령구분은 이 데이터에서 가져온다.
- 제이스 C (부서이동 + 기본권한): 발령구분=부서이동 AND 권한행=일반사용자
→ 회수 대상 없음, 자동 Pass.

[판정 기준 공통]

- 확인필요: 회수 이력 데이터 부재, 중복 누락, 데이터 형식 오류로 판단 불가 시
 - 1. 제이스 C의 성립 조건이 데이터상 충족될 수 있다. 회수 대상 권한이 없는 발령자는 회수 이력에 기록이 생성되지 않으므로 권한 행을 기록할 수 없다. '회수 대상이 없는 사항'을 '회수 기록에서 위은 권한행이 일반사용자인 사항'으로 정의한 것이다.
 - 2. 두 구간이 같은 행에 다른 결론을 지시한다. 제이스 C는 Pass를 공통 조항은 확인필요를 지시하는데 우선순위가 명시되어 있지 않다.
 - 3. 대상 시스템의 처리 단계가 프로그래밍에 없다. 부서이동 시 추가 부여 권한만 자동 회수한다는 사실이 전행 9504자 이데에도 없다(추가 부여: "회수하지" 회수 대상이 아* 모두 0 회). 인간 평가자는 이 편차를 알고 PASS로 판정하였다.
- 실적 확인: 인사발령 20명(부서이동 16회지 4) 중 권한 회수 이력에 등장하는 사람은 4명이며 전행 퇴직자이고, 부서이동자 중 회수 이력 보유자는 0명이다.

F-3. 보안 조건과 결과

저장된 웹 프로그래밍에 문장을 삽입한 하고 나머지는 바이트 단위로 동일하게 유지하였다. 재현성이 관심사이므로 각 조건을 3회 반복하였다.

P1: 시스템 전체를 (발령 선택리스트에 추가 (+120자)
시스템 전체 부서이동 발생 시 추가 부여 권한(ROLE_USER)이 있는 경우에만 해당 권한을 자동 회수하며, 일반사용자(ROLE_USER) 권한만 보유한 경우 회수 대상이 없으므로 회수 이력이 생성되지 않는다.

P2: P1에 대해 편향 기준 공통에 우선순위를 명시 (+252자)
- 단, 위 시스템 전체에 따라 회수 대상 권한이 존재하지 않는 경우는 '회수 이력 데이터 부재'에 해당하지 않는다. 부서이동 발령자가 권한 회수 이력에 없으면 추가 부여 권한이 없는 것이므로 제이스 C로 분류하고 Pass로 판정한다.

조건	가용할 데이터 일치	37건 구간의 편향	회차 간 불일치
2026-05 운영 실행	50/50	PASS 37	-
보완 전 (C4-C4+2, C4+3-C9)	29/50 (4회 모두)	PASS 16 고정	최대 15행
P1 (전체판)	35/50 - 50/50 - 50/50	PASS 22~37	1회차만 이합 (15행)
P2 (전체 + 우선순위)	50/50 - 50/50 - 50/50	PASS 37	0

P2의 3회 실행은 행 단위로 완전히 일치하였으며, 2026년 5월 실행 결과에도 동일하였다. P1의 23회차도 P2와 한 행의 차이 없이 같았다(6회 중 5회 완전 일치).
보완 전 조건은 4회 실행에서 한 번도 37건 구간의 PASS가 16건을 넘지 못하였다. 편차를 추가회차 3회 중 2회까지 37건 전전으로, 우선순위를 명시하자 3회 전구간 37건 전전으로 수렴하였다.

F-4. 편향의 원인

동일 중재에 대한 동일 편향은 모형을 고장하는 것만으로 확보되지 않으며, 편향 기준이 대상 시스템의 처리 논리를 빠짐없이 포괄할 때 확보된다. 다른 이유를 살펴본다.

- 분포이 보고한 수치는 보완 이전의 프로그래밍에 의거한 값이며, 본 절의 결과로 대체하지 않았다. P1-P2는 원인 규명을 위한 추가 실험이다.
- 대상은 한 통재(계정-관련 회수)이며, 다른 통재나 다른 시스템으로의 일반행에는 추가 검증이 필요하다.
- 고장능 모델은 편차가 지정되지 않은 범정으로 호출되므로, 재현을 시도하는 경우 날짜가 지정된 스냅샷을 병행하여 사용될 것을 권고한다.

F-5. 재현

```
# 보완 조건 실행 (P1 또는 P2, 각 3회 반복)
PYTHONPATH=$PWD uv run python -m experiments.premise P2 experiments/results

# 무편향 조건 4회 실행의 행 단위 실행 비교 (F-1 표)
PYTHONPATH=$PWD uv run python -c "
import json, itertools, collections
from pathlib import Path
```

```
R=Path('experiments/results')
KEY=(IT02.R05.C01:'배우번호', IT01.R01.C01:'사번', IT01.R02.C01:'발령자_사번')
data={}
for c in ['C4', 'C4+2', 'C4+3', 'C9']:
    d={}
    for ctk in KEY.items():
        for r in json.loads(R/rows_{c}_{ctk}.json).read_text(encoding='utf-8'):
            d[ctk]=r['fields']['overall result']
    data[c]=d
for a,b in itertools.combinations(data):
    common=set(data[a]&data[b]).diff=[x for x in common if data[a][x]!=data[b][x]]
    print(a,b,len(common),len(diff),dict(collections.Counter(x[0] for x in diff)))
```

실행 위치와 문장은 experiments/premise.py의 PREMISE-PRIORITY 상수에 정의되어 있으며, 실행 시 실행 외의 변경이 없음을 일기로 검증한다.

G. 가상 중재의 생성 절차

실사화된 1-3이 요청한 생성 절차: 생성에 사용된 모델, 설계자의 문제, 개발/평가 데이터의 분리 기술을 기술한다.

생성 주제와 역할 분담	수행 주제
가상 HR 시스템(유형 산출 시스템)	평가 시스템과 설계의 프로그래밍 개발
화면-문서 형식	연구자가 구성을 제시하고 생성형 AI가 구현
시스템 데이터 생성 코드	Anthropic Claude Code 환경의 에이전트 팀 — claude-opus-4(중형), claude-sonnet-4-6(개발 테스트), claude-hiku-4-5-2025(100)(단순 작업)
시나리오(형식-구조-미숙인 등의 대상 사례로 조직)	생성 과정에서 생성형 AI가 결정
업무 처리 형식(예: 청상 문장은 승인 후 바로, 긴급 발행은 전배로 후 사용 승인)	연구자가 제시
결함 사례의 지정	연구자가 특정 사례를 일함으로 지정하지 않음. 개별 워크로드의 내용과 결함 행의 여부는 검토하지 않음

판단 규칙과 데이터의 관계

- 판단 규칙의 출처는 가상 실무의 IT 일반용제 위험통제매트릭스(RCM)이며, 매트릭스는 2026년 3월 30일에 확정되어 중형 987개는 미지의 유형에 대해 미확화 성능이 아니라 판단 규칙이 해당 유형에 적합한 상태에서의 일함이다(판문 N/A형 일함화 형제). 다만 보류된 가상 일함화 성능은 중형에 대한 편향 부기의 추가와 공통 오츠의 수위에 확정되었고, 기준과 평가자가 알고 있던 대상 시스템의 처리 단계가 프로그래밍에 포함되지 않아 편향이 기준값과 판문 구간이 재실행에서 판별되었다(F 표). 설계 요소의 기여는 판단 규칙을 고장한 조건 간 상대 비교로 측정하였다(테블, 본문 IV 장 2 절).

일정

사건	사건
2026-03-30	판단 규칙의 출처(RCM) 확정
2026-04 ~ 05-07	가상 HR 시스템 개발 및 중형 데이터 생성, 모집단 확정
2026-05-06 ~ 05-10 이전	기준값 조사(숙인 평가자) 작성 — 시스템 평가 실행 전 완료
2026-05-07 ~ 05-11	판단 규칙(중형별 세부 지시사항) 보완
2026-05-10 ~ 05-13	시스템 평가 실행 (논문 보고 수지)

개발/평가 데이터의 분리

개발/평가와 평가용 중형을 분리하지 않았으며, 판단 규칙의 보완 과정에서 평가 대상 중형의 실행 결과를 참조하였다. 따라서 본문의 987개는 미지의 유형에 대해 미확화 성능이 아니라 판단 규칙이 해당 유형에 적합한 상태에서의 일함이다(판문 N/A형 일함화 형제). 다만 보류된 가상 일함화 성능은 중형에 대한 편향 부기의 추가와 공통 오츠의 수위에 확정되었고, 기준과 평가자가 알고 있던 대상 시스템의 처리 단계가 프로그래밍에 포함되지 않아 편향이 기준값과 판문 구간이 재실행에서 판별되었다(F 표). 설계 요소의 기여는 판단 규칙을 고장한 조건 간 상대 비교로 측정하였다(테블, 본문 IV 장 2 절).