Korean Journal of Clinical Dental Hygiene

Original Article

치과위생사 국가시험에서 생성형 AI의 프롬프트 전략별 정답률과 오류 패턴

김상겸*
남서울대학교 치위생학과

Accuracy and Error Patterns of Generative AI on the Korean National Dental Hygienist Licensing Examination Across Prompt Engineering Strategies

Received: June 19, 2026
Revised: July 15, 2026
Accepted: August 04, 2026

Sang-Kyeom Kim*
Department of Dental Hygiene, Namseoul University, Cheonan, 31020, Korea

*Corresponding Author: 김상겸, (31020) 충청남도 천안시 서북구 성환읍 대학로 91 남서울대학교 치위생학과, Tel: (041) 580-2564, Fax: (041) 580-2927, E-mail: kyeoms19@nsu.ac.kr

Abstract

Objective: This study aimed to compare GPT-4o accuracy across five prompt engineering strategies using items from the Korean National Dental Hygienist Licensing Examination and to examine differences by question domain and type. **Methods:** A total of 399 items from the 2024-2025 examinations were classified by domain (Basic/Clinical) and type (Recall/Reasoning) and presented to GPT-4o (GPT-4o-2024-11-20, temperature=0) under five prompt conditions. Differences in accuracy across prompt conditions were assessed using Cochran's Q test. For reasoning-type items, exploratory pairwise comparisons were performed using exact McNemar tests with Holm correction. **Results:** Overall accuracy ranged from 75.2% to 78.9%, exceeding the overall threshold of 60% under all prompt conditions. Cochran's Q showed no significant overall difference among the five prompt conditions (Q=6.462, p=0.167). For reasoning-type items (n=80), the combined prompt yielded the highest accuracy(75.0%) compared with 65.0%–67.5% for the other conditions. However, the overall difference among prompt conditions was not statistically significant (Q=6.759, p=0.149), and no pairwise comparison remained significant after Holm correction. Qualitative analysis identified recurring hallucinations involving Korea-specific regulations. **Conclusions:** GPT-4o achieved accuracy exceeding the 60% overall threshold regardless of prompt strategy, suggesting that the overall effect of prompt engineering was limited. Although the combined prompt showed a favorable trend for reasoning-type items, statistical evidence for this effect was limited. These findings highlight the importance of AI literacy for the critical evaluation of AI-generated rationales.

Korean Journal of Clinical Dental Hygiene, 2026;14(3):209-219 https://doi.org/10.12972/kjcdh.2026.14.3.4 pISSN : 2287-9641 eISSN : 2465-7743 Copyright © 2026 by Korean Journal of Clinical Dental Hygiene. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License(https://creativecommons.org/licenses/by-nc/4.0/).

Key words: Artificial intelligence, Dental hygiene education, Dental hygienists, Large language models, Prompt engineering

색 인: 인공지능, 치위생교육, 치과위생사, 거대언어모델, 프롬프트 엔지니어링

1. 서론

최근 ChatGPT, GPT-4 등 대형언어모델(Large Language Model, LLM)의 발전과 더불어, 의료 및 보건의료 분야에서 LLM의 활용 가능성에 관한 연구가 활발히 이루어지고 있다. 보건의료 교육, 연구, 실무 전반에서 생성형 AI 도입에 대한 기대와 우려가 보고되어 왔다(1, 2). 특히 각국의 보건의료인 국가시험 문항을 LLM에 입력하여 정답률을 평가하는 연구들이 의학, 치의학, 간호학, 약학 등 다양한 분야에서 보고되고 있으며, 이는 LLM이 해당 직역의 전문지식을 어느 정도 수준으로 보유하고 있는지를 가늠하는 척도로 활용되고 있다.

치과위생사 국가시험을 대상으로 한 연구도 점차 보고되고 있다. Bae SM 등(3)은 ChatGPT-3.5에 제49회 치과위생사 국가시험 200문항을 입력하여 정답률 45.5%를 확인하였으며, 오답 102문항 중 100문항이 환각(Hallucination)에 해당함을 보고하였다. Song ES and Lee SP(4)는 Gemini, GPT-3.5, GPT-4를 2019-2023년 5개년 기출문항(한국어 및 영어)에 적용하여 모델 간 정답률을 비교하였으며, GPT-4가 평균 65.9%로 가장 높았으나 치위생(학)과 학생의 평균 합격률(81.9%)에는 미달하였다. 이 두 연구는 모두 단일한 입력 방식(Zero-shot)으로 모델의 정답률을 측정하는 데 초점이 맞추어져 있었다.

한편 치과의사 국가시험(Korean Dental Licensing Examination, KDLE)에서는 Kim W et al.(5)이 GPT-3.5, GPT-4, Claude3-Opus를 비교하여 GPT-4와 Claude3-Opus의 합격선 통과를 보고하였고, Song ES et al.(6)은 비교적 최신 모델인 GPT-4o가 Gemini Advanced보다 유의하게 높은 정답률(81.1% vs. 76.6%)을 보임을 확인하였다. 치과위생사 국가시험에서는 Choi EM과 Park SK(7)이 2023년 기출문항 200문항을 활용하여 GPT-4o와 GPT-o1의 성능을 비교하였으며, 각각 77.5점과 75.5점으로 두 모델 간 차이는 통계적으로 유의하지 않았다고 보고하였다. 그러나 해당 연구들은 모델 간 단순 성능 비교에 초점을 두었으며, 프롬프트 전략에 따른 정답률 변화, 문항 유형별 반응 양상, 오류 패턴은 함께 분석하지 않았다. 따라서 치과위생사 국가시험 문항을 활용한 LLM의 프롬프트 전략별 성능과 오류 양상을 체계적으로 평가한 연구는 여전히 제한적이다. 본 연구는 단일 입력 방식에 그친 선행연구와 달리, 동일 문항에 대해 5가지 프롬프트 전략을 체계적으로 적용 및 비교하고 문항 특성(영역×유형)에 따른 반응 차이와 반복적 오류 패턴을 함께 분석함으로써, 치과위생사 국가시험에서의 생성형 AI 활용에 관한 보다 체계적인 근거를 제공하고자 한다.

프롬프트 엔지니어링(Prompt engineering)은 모델의 가중치를 변경하지 않고도 입력 방식만으로 LLM의 응답 품질을 개선할 수 있는 방법으로, Few-shot learning, 역할 부여(Role prompting), 사고연쇄(Chain-of-Thought, CoT) 등이 대표적이다(8-10). 다른 한국 보건의료 국가시험을 대상으로 한 연구에서 그 효과가 시사된 바 있는데, Jang D et al.(11)은 한의사 국가시험에서 한자주석, 영어 번역, 시험맞춤형 지시문 등을 단계적으로 적용하여 GPT-4의 정답률을 51.82%에서 66.18%로 크게 향상시켰다. 반면 Hsieh MY et al.(12)은 의과대학 시험에서 GPT-4o 등 고성능 모델에서는 프롬프트 엔지니어링의 효과가 미미해지는 경향을 보고하였다. 의학교육 전반에서 LLM 활용 동향을 정리한 Xu X et al.(2)도 프롬프트 활용 능력 함양의 중요성을 강조하였으나, 이를 실제 보건의료인 국가시험에 체계적으로 적용, 검증한 연구는 제한적이다(13).

이에 본 연구는 2024년 및 2025년 치과위생사 국가시험 기출문항을 대상으로, GPT-4o에 대해 다섯 가지 프롬프트 전략(Zero-shot, Role prompting, 간결화된 Chain-of-Thought, Few-shot, 통합형)을 적용하여 정답률을 비교하고자 한다. 또한 문항을 영역(기초/임상) 및 유형(암기형/추론형)에 따라 분류하여, 프롬프트 전략의 효과가 문항 특성에 따라 달라지는지를 함께 분석함으로써, 치과위생사 및 치위생(학)과 교육 현장에서 생성형 AI를 보다 효과적으로 활용하기 위한 실용적 지침을 제시하고자 한다.

2. 연구대상 및 방법

2.1. 연구설계 및 자료

본 연구는 2024년 및 2025년에 시행된 치과위생사 국가시험 기출문항을 대상으로, 동일 문항에 대해 서로 다른 프롬프트 전략을 적용하였을 때 GPT-4o의 정답률 차이를 비교하는 반복측정 비교 연구(Repeated-measures comparative study)이다.

분석 대상 문항은 2024년(1교시, 2교시) 및 2025년(1교시, 2교시) 기출문항 중 그림 또는 사진 자료 공개가 제한되어 풀이가 불가능한 문항 1개를 제외한 총 399문항이었다. 5개 프롬프트 조건을 각 문항에 적용하여 총 1,995회(399문항×5조건)의 응답을 수집하였다.

2023년 기출문항은 분석 대상에서 제외하고, Few-shot 프롬프트(P4, P5)에 사용되는 예시 문항으로만 활용하였다. 이는 평가 문항(2024–2025)과 예시 문항(2023)을 분리함으로써 실험 내 직접적인 문항 누출을 최소화하고자 한 것이다. Few-shot 예시는 영역(기초/임상)×유형(암기형/추론형) 조합별로 1문항씩, 총 4문항을 1교시, 2교시에서 균형 있게 선정하였다.

2.2. 평가모델 및 실험환경

평가 모델로는 GPT-4o(버전: GPT-4o-2024-11-20)를 사용하였다. 모든 응답은 OpenAI API를 통해 수집하였으며, 응답의 재현성을 확보하기 위해 Temperature는 0으로 고정하였다.

2.3. 프롬프트 조건

본 연구에서는 임상 현장에서 치과위생사 또는 치위생(학)과 학생이 생성형 AI를 활용할 때 나타날 수 있는 패턴을 반영하여 다음과 같은 5가지 프롬프트 조건을 설정하였다(Table 1).

P1(Zero-shot)은 문항과 선택지를 그대로 제시하고 정답 번호만 숫자로 답하도록 요청하는 조건으로, 가장 일반적인 복사-붙여넣기 사용 패턴을 반영한다(8). P2(Role prompting)는 P1에 '당신은 치과위생사 국가시험 전문가입니다'라는 역할 지시를 추가한 조건이다(9). P3(간결화된 Chain-of-Thought, CoT)는 간단한 근거를 바탕으로 정답을 고르도록 지시하고, 자동 채점을 위해 '정답: [숫자]' 형식으로 응답하도록 표준화한 조건이다. 본 연구에서는 자동화된 정답 추출의 신뢰성을 확보하기 위해 전형적인 단계별 추론 형식(10)을 간결화하였으며, 이는 의료 국가시험 분야 선행연구에서도 채택된 방식이다(13, 14). P4(Few-shot)는 2023년 기출문항 중 영역×유형 조합별 1문항씩 총 4문항을 정답과 함께 예시로 제시한 후 본 문항을 질의하는 조건이다. P5(통합형)는 P2(역할 부여), P3(근거 요청), P4(Few-shot 예시)를 모두 결합한 조건이다.

P1에서 P5로 갈수록 단순한 사용에서 체계적인 사용으로 이어지는 단계적 구성을 의도하였으며, 각 조건은 누적적으로 요소를 추가하는 구조를 갖는다. 단, P5의 효과는 개별 요소의 단순 합이 아닌 결합 효과로 해석하였다.

Table 1. Description of the five prompt engineering strategies applied to GPT-4o

Condition	Strategy	Description
P1	Zero-shot	Question and answer choices presented as-is, with the instruction to respond with the answer number only
P2	Role prompting	P1 plus the instruction "You are an expert in the Korean National Dental Hygienist Licensing Examination"
P3	Simplified CoT	Instruction to provide a brief rationale before selecting the answer, with responses standardized in the format "Answer: [number]"
P4	Few-shot	Four example questions (with correct answers) from the 2023 examination, covering all combinations of domain (basic/clinical) and question type (recall/reasoning), presented before the target question
P5	Combined (Role + CoT + Few-shot)	Integration of the role assignment (P2), rationale-based response (P3), and few-shot examples (P4)

GPT-4o, GPT-4o-2024-11-20; P1, Zero-shot; P2, Role prompting; P3, Simplified Chain-of-Thought; P4, Few-shot; P5, Combined (Role + Chain-of-Thought + Few-shot); CoT, Chain-of-Thought.

2.4. 문항 분류

치과위생사 국가시험은 의료관계법규, 치위생학1(기초치위생, 치위생관리), 치위생학2(임상치위생처치, 임상치과지원)의 총 21개 과목, 200문항(1교시, 2교시 각각 100문항)으로 구성된다(15). 본 연구에서는 문항 내용을 바탕으로 연구자 1인이 아래 기준에 따라 영역(Domain) 및 유형(Question type)으로 분류하였다.

영역은 해부학·생리학·병리학·미생물학적 지식, 법규, 통계, 행정, 교육 이론 등 환자나 임상 상황과 무관한 일반적 지식, 개념, 원리를 묻는 문항을 '기초'로, 특정 임상 상황에서의 처치, 시술, 기구, 재료 등의 적용을 묻는 문항을 '임상'으로 정의하였다. 유형은 단일 사실, 정의, 분류 체계에 대한 직접적인 회상을 요구하는 문항을 '암기형'으로, 임상 상황 및 사례가 제시되고 이에 대한 적용, 판단, 추론을 요구하는 문항을 '추론형'으로 정의하였다.

신뢰도 검증을 위해 치위생학 전공 추가 평가자 1인이 층화 무작위 표본추출로 선정된 40문항을 동일한 기준에 따라 독립적으로 분류하였다. 초기 분류 결과에서 불일치한 문항에 대해서는 두 평가자가 분류 기준을 재검토하고 합의를 통해 최종 분류를 결정하였다. 합의 후 평가자 간 신뢰도는 영역 분류 κ=0.784, 유형 분류 κ=0.778로 'Substantial' 수준의 일치도를 보였다.

2.5. 통계분석

5개 프롬프트 조건 간 정답 여부의 차이는 동일 문항에 대해 반복 측정된 이분형(Binary) 자료임을 고려하여 Cochran's Q test로 검정하였다. 전체 문항 및 문항 유형(암기형/추론형)별로 각각 분석을 수행하였다. 추론형 문항에 대해서는 P5와 나머지 조건(P1-P4) 간 쌍별 비교를 Exact McNemar test를 이용하여 탐색적으로 수행하였으며, 다중비교에 따른 제1종 오류를 보정하기 위해 Holm correction을 적용하였다. 영역×유형 조합(기초-암기형, 기초-추론형, 임상-암기형, 임상-추론형)에 따른 분석은 탐색적(Exploratory) 분석으로 별도 제시하였다.

모든 통계분석은 Python (version 3.12.8, Python Software Foundation, Wilmington, DE, USA)을 이용하여 수행하였다. Cochran's Q test는 표준 공식을 이용하여 직접 산출하였고, exact McNemar test의 p값은 Python의 SciPy 라이브러리 (version 1.13.1, SciPy Contributors, open-source)의 Binomial exact test를 이용하여 계산하였다.

2.6. 윤리적 고려사항

본 연구는 사람 또는 동물을 대상으로 한 연구가 아니므로 기관생명윤리위원회(Institutional Review Board, IRB)의 심의 대상에 해당하지 않는다.

본 연구에서 사용한 국가시험 문항의 원문 및 선택지는 저작권 보호 대상일 수 있음을 고려하여, 본문, 표, 부록 및 공개 자료에 수록하지 않았다. 연구 결과는 문항을 식별하기 위한 메타데이터(시행연도, 교시, 문항번호, 영역, 유형) 및 집계된 분석 결과를 중심으로 제시하였다.

3. 연구결과

3.1. 전체 정답률 비교

전체 399문항에 대한 5개 프롬프트 조건별 정답률은 P1 75.4%(301/399), P2 75.2%(300/399), P3 76.7%(306/399), P4 76.2%(304/399), P5 78.9%(315/399)였다(Table 2). Cochran's Q test 결과 5개 조건 간 전체 정답률 차이는 통계적으로 유의하지 않았다($Q=6.462$, $p=0.167$). 다만 가장 체계적인 프롬프트 조건인 P5에서 최고 정답률을 보였으며, 모든 조건의 정답률이 치과위생사 국가시험의 전체 총점 기준 합격선인 60%를 상회하였다.

Table 2. Accuracy of GPT-4o by prompt condition and Cochran's Q test results (N = 399)

Prompt condition	Correct (n)	Accuracy (%)
P1	301	75.4
P2	300	75.2
P3	306	76.7
P4	304	76.2
P5	315	78.9

Cochran's Q test: $Q=6.462$, $p=0.167$. P1, Zero-shot; P2, Role prompting; P3, Simplified Chain-of-Thought; P4, Few-shot; P5, Combined (Role + Chain-of-Thought + Few-shot).

3.2. 문항 유형별 분석

문항을 암기형(n=319)과 추론형(n=80)으로 구분하여 분석한 결과, 두 유형 간 프롬프트 전략에 대한 반응 양상은 다소 차이를 보였다(Figure 1).

암기형 문항에서는 프롬프트 조건 간 정답률이 77.4%(P1)–79.9%(P5)로 비교적 일정하게 유지되었으며, Cochran's Q test 결과 통계적으로 유의한 차이는 없었다($Q=3.018$, $p=0.555$).

추론형 문항에서는 P1–P4에서 65.0–67.5%의 정답률을 보인 반면, P5에서는 75.0%로 가장 높은 정답률을 보였다. 그러나 Cochran's Q test에서는 통계적으로 유의한 차이가 확인되지 않았다($Q=6.759$, $p=0.149$). P5와 나머지 조건 간 쌍별 비교를 exact McNemar test로 탐색적으로 수행한 결과, P5 vs. P4 비교에서 가장 작은 Unadjusted p값($p=0.039$)을 보였으나, Holm correction 적용 후에는 모든 쌍별 비교에서 통계적 유의성이 유지되지 않았다(Adjusted $p=0.154$–0.292).

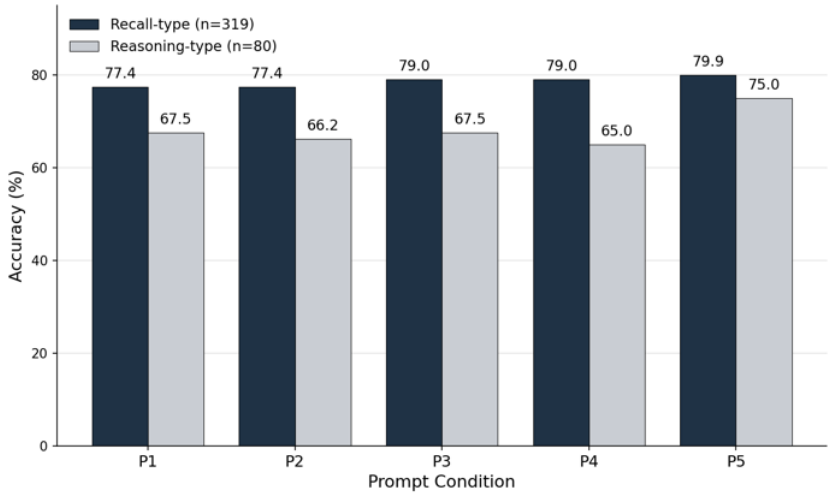


Figure 1. Comparison of GPT-4o accuracy by question type (Recall vs. Reasoning) across the five prompt conditions (P1–P5). P1, Zero-shot; P2, Role prompting; P3, Simplified chain-of-thought; P4, Few-shot; P5, combined (Role + Chain-of-Thought + Few-shot).

3.3. 영역×유형 조합별 분석(탐색적 분석)

문항을 영역(기초/임상)과 유형(암기형/추론형)의 조합으로 세분화하여 탐색적으로 분석한 결과, 4개 그룹(기초-암기형 n=229, 기초-추론형 n=25, 임상-암기형 n=90, 임상-추론형 n=55) 모두에서 Cochran's Q test 결과 프롬프트 조건 간 통계적으로 유의한 차이는 확인되지 않았다(각 p=0.682, 0.683, 0.226, 0.187)(Figure 2).

다만 임상-추론형(n=55) 그룹에서 P1–P4(65.5–67.3%) 대비 P5(76.4%)가 약 9–11%p 높은 정답률을 보여, 전체 추론형 분석에서 관찰된 P5의 우위가 주로 임상-추론형 문항에서 기인하였을 가능성을 시사한다. 반면 임상-암기형(n=90) 그룹에서는 P1(76.7%)에서 P5(84.4%)로 갈수록 점진적으로 정답률이 상승하는 경향을 보였다. 기초-암기형 및 기초-추론형 그룹에서는 프롬프트 조건 간 뚜렷한 경향성은 관찰되지 않았다.

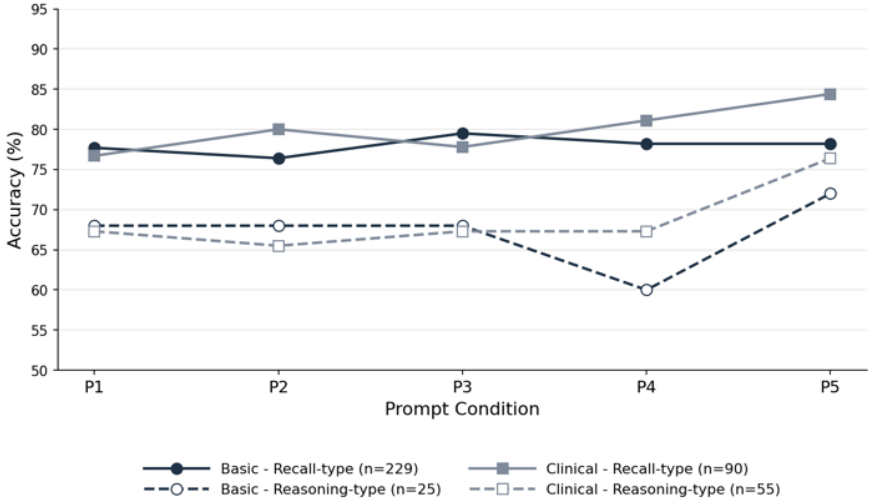


Figure 2. Accuracy trends by domain×question type subgroup across the five prompt conditions (P1–P5) (Exploratory). P1, Zero-shot; P2, Role prompting; P3, Simplified chain-of-thought; P4, Few-shot; P5, combined (Role + Chain-of-Thought + Few-shot).

3.4. 연도별 비교 분석

2024년 기출문항(n=200)과 2025년 기출문항(n=199)의 정답률을 비교한 결과, 모든 프롬프트 조건에서 2024년 문항의 정답률이 2025년 문항보다 높았다(P1: 79.5% vs. 71.4%, 격차 8.1% p; P2: 81.5% vs. 68.8%, 격차 12.7% p; P3: 80.0% vs. 73.4%, 격차 6.6% p; P4: 80.0% vs. 72.4%, 격차 7.6% p; P5: 81.0% vs. 76.9%, 격차 4.1%p). 이러한 연도별 차이에는 문항 난이도, 문항 구성, 유형별 분포, 공개 자료 노출 가능성, 모델의 학습 데이터 시점 등 여러 요인이 복합적으로 작용했을 가능성이 있으며, 본 연구의 설계로는 특정 요인의 영향을 분리하여 확인하기 어렵다. 다만 5개 조건 중 P5에서 연도 간 격차가 가장 작게 나타난 점은 주목할 만하다.

3.5. 오답 패턴 분석

오답 사례에 대한 질적 검토 결과, 반복적으로 관찰되는 오류 패턴을 4개 유형으로 정리하였다(Table 3). 첫째, 지역 법규/정책 환각(Local legal/Policy hallucination)의 대표적 예로, 2024년 1교시 17번 문항(「구강보건법」상 국민 구강건강실태조사·공표 의무자, 정답: 질병관리청장)에서 5개 프롬프트 조건 모두 동일하게 '보건복지부장관'을 정답으로 제시하였다. 둘째, 통계지표 개념 혼동(Rate vs. Index)도 반복적으로 관찰되었다. 2024년 1교시 87번 및 2025년 1교시 85번 문항에서, GPT-4o는 인구 기반 유병률(Rate) 산출 공식을 적용해야 하는 문항에 대해 치아 기반 지수(Index) 산출 공식을 적용하는 동일한 패턴의 오류를 보였다. 셋째, 2024년 1교시 30번 문항(복근치/이근치 분류)에서는 해부학적 세부지식 오류를 보였다. 넷째, 2025년 2교시 10번 문항에서는 회전법(Rolling method)이 정답인 문항에 대해 스틸만법(Stillman method)을 정답으로 제시하는 임상 기준 불일치(Clinical guideline mismatch)를 보였다.

한편, 첫째와 넷째 유형의 사례에서는 공통적으로 그럴듯하지만 근거가 부정확한 설명(Plausible but unsupported rationale)이 함께 생성되는 경향이 관찰되었다. 17번 문항에서 P3와 P5는 각각 「구강보건법」 제12조 및 제16조를 근거로 제시하였는데, 서로 다른 조문 번호를 인용하면서도 동일한(오답) 결론에 도달하였다. 10번 문항에서도 GPT-4o는 스틸만법의 적응증에 대해 표면적으로 그럴듯하지만 교과서 기준과는 일치하지 않는 근거를 함께 제시하였다. 이는 CoT 기반 프롬프트가 형식적으로 근거를 제시하더라도 그 근거 자체의 사실성은 보장되지 않으며, 오히려 오답에 대한 확신을 강화하는 형태로 나타날 수 있음을 보여준다.

Table 3. Typology of recurring error patterns identified through qualitative analysis

Error type	Description	Example item(s)
Local legal/policy hallucination	Generating factually incorrect information about Korea-specific regulations or the responsible administrative authority	2024-S1-17
Conceptual confusion (rate vs. index)	Confusing population-based epidemiological rates with tooth-based indices, leading to incorrect formula application	2024-S1-87, 2025-S1-85
Anatomical/detail knowledge error	Incorrect classification of detailed anatomical terminology	2024-S1-30
Clinical guideline mismatch	Selecting a clinical procedure inconsistent with standard textbook criteria	2025-S2-10

S1/S2, Session 1/Session 2 of the examination.

4. 고안

본 연구의 결과는 두 가지 핵심 메시지로 요약될 수 있다. 첫째, 암기형 문항에서는 프롬프트 전략과 무관하게 GPT-4o가 안정적인 정답률(77.4–79.9%)을 보였다. 이는 단순한 사실 확인이나 정의를 묻는 질문의 경우, 별도의 프롬프트 기법을 적용하지 않고 문항을 그대로 입력하는 가장 단순한 방식(P1)만으로도 일정 수준 활용 가능한 정답률을 얻을 수 있음을 의미한다. 둘째, 추론형 문항에서는 통합 프롬프트(P5)가 다른 조건들보다 수치상 가장 높은 정답률(75.0%)을 보였으며, 이러한 경향은 특

히 임상적 판단을 요하는 임상-추론형 문항(76.4%)에서 두드러졌다. 다만 Cochran's Q test 결과 전체 조건 간 차이는 유의하지 않았고(p=0.149), 탐색적 McNemar 쌍별 비교에서도 Holm 보정 후 통계적 유의성은 확인되지 않았다(P5 vs. P4 adjusted p=0.154). 따라서 이 결과는 확정적 효과라기보다, 향후 더 큰 표본에서 검증이 필요한 경향으로 해석하는 것이 적절하다.

치과위생사 국가시험의 필기시험 합격 기준은 전 과목 총점의 60% 이상 및 매 과목 만점의 40% 이상 득점이다(16). 본 연구에서 관찰된 전체 정답률(75.2-78.9%)은 모든 프롬프트 조건에서 이 기준을 상당한 폭으로 상회하였다. 이는 GPT-4o가 별도의 정교한 프롬프트 설계 없이도, 즉 단순한 질의(P1) 수준만으로도 치과위생사 국가시험의 전체 총점 기준 합격선을 상회하는 정답률을 달성하였음을 보여준다. 다만 실제 치과위생사 국가시험의 합격 여부는 전체 총점 기준(60% 이상)뿐 아니라 매 과목 만점의 40% 이상이라는 과목별 과락 기준도 동시에 충족해야 하며, 본 연구는 21개 과목 단위의 정답률 및 과락 기준 충족 여부를 별도로 분석하지 않았다. 따라서 본 연구에서 관찰된 정답률이 합격선을 상회한다는 결과를 실제 국가시험 합격과 동일하게 해석하는 것은 적절하지 않다.

본 연구에서 설정한 5가지 프롬프트 조건은 실제 치과위생사 또는 치위생(학)과 학생이 생성형 AI를 사용하는 다양한 패턴을 반영한 것이다. 본 연구 결과에 따르면, 단순 사실 확인 목적이라면 P1과 같은 단순한 사용만으로도 충분한 반면, 임상적 판단이 요구되는 상황에서는 P5와 같이 역할을 부여하고 근거를 요청하며 예시를 함께 제시하는 방식이 상대적으로 안정적인 응답을 얻는 데 도움이 될 가능성이 있다. 다만 P4와 P5는 Few-shot 예시를 사전에 준비해야 한다는 점에서 일반 사용자에게는 진입장벽이 있을 수 있다. 이러한 점에서 P2(역할 부여)나 P3(근거 요청)와 같이 추가 준비 부담이 적으면서도 일정 수준의 효과를 기대할 수 있는 전략이 현실적인 대안으로 고려될 수 있다.

본 연구의 P1(Zero-shot) 정답률(75.4%)은 동일 시험을 ChatGPT-3.5로 평가한 Bae SM 등(3)의 결과(45.5%) 및 GPT-4로 평가한 Song ES and Lee SP(4)의 결과(65.9%)보다 높았으며, 이는 모델 세대(GPT-3.5에서 GPT-4, GPT-4o로 이어지는)에 따른 향상 추세와 일치한다. 동일 모델(GPT-4o)을 적용한 국가시험에서는 일본 의사 국가시험에서 93.2%(17), 일본 치과의사 국가시험에서 73.8%(18), 한국 치과의사 국가시험에서 81.1%(6)가 보고된 바 있어, 본 연구의 정답률(75.2-78.9%)은 그 사이에 위치한다. 이는 직군, 언어, 도메인별 난이도 차이를 반영할 수 있으며, 직군 간 비교를 위한 후속 연구가 필요함을 시사한다. 한편 본 연구에서 전체 정답률 측면의 프롬프트 조건 간 차이가 통계적으로 유의하지 않았던 것은, 이미 높은 성능을 보이는 최신 모델에서는 프롬프트 엔지니어링의 전반적 효과가 감소한다는 Hsieh MY et al.(12)의 보고와 일치한다. 본 연구에서는 추론형 문항이라는 특정 하위집단에서 통합 프롬프트(P5)가 다른 조건들보다 수치상 가장 높은 정답률을 보이는 경향이 관찰되었으며, 이는 여러 프롬프트 기법의 결합이 효과적일 수 있음을 시사한 선행 연구 결과(8, 10, 11, 19)와 같은 방향성을 보인다. 다만 통계적 유의성에는 이르지 못하였으므로, 이러한 경향이 실제 효과를 반영하는지 여부는 향후 더 큰 표본을 통한 검증이 필요하다. 아울러 모든 프롬프트 조건에서 2024년 문항 대비 2025년 문항의 정답률이 낮게 나타난 데에는 문항 난이도, 문항 구성, 공개 자료 노출 가능성, 모델 학습 데이터 시점 등 여러 요인이 복합적으로 작용했을 가능성이 있으며, 이러한 연도별 격차가 P5에서 가장 작게 나타난 점(4.1%p)은 후속 연구를 통한 검증이 필요한 흥미로운 경향으로 판단된다.

질적으로 검토한 오답 사례들은(Table 3) 한국 특화 법규나 통계지표와 같이 국가 및 지역 특이적 지식이 요구되는 영역에서 환각(Hallucination) 또는 개념 혼동이 반복적으로 나타나며, 임상적 판단을 요하는 문항에서는 표면적으로 그럴듯하지만 실제 교과서 기준과는 일치하지 않는 근거를 제시하는 경우가 있음을 보여주었다. 이러한 오류들은 사실과 다르지만 그럴듯한 내용을 생성하는 환각의 전형적 양상으로 해석할 수 있으며(20, 21), 의료 분야에서 환각은 임상적으로 중요한 오류로 이어질 수 있다. 이러한 결과는 LLM이 제시하는 답변, 특히 그 근거(Rationale)를 비판적으로 검토하지 않고 그대로 수용할 경우 오개념이 형성될 위험이 있음을 시사한다. 따라서 치위생(학)과 교육과정에서는 생성형 AI를 단순히 정답을 제공하는 도구로 활용하는 것을 넘어, LLM이 제시한 근거의 타당성을 학습자가 스스로 검토하고 판단할 수 있는 능력, 즉 AI 리터러시(AI literacy)를 함양하는 교육이 병행될 필요가 있다(22, 23). 본 연구에서 확인된 구체적인 오류 유형(Table 3)들은 이러한 AI 리터러시 교육의 사례 자료로 활용될 수 있을 것이다. 구체적으로는, 한국 특화 법규 영역의 환각 사례는 AI 응답의 무비

판적 수용 위험성을 교육하는 자료로, 개념 혼동 및 임상판단 오류 사례는 학습자가 AI 생성 근거의 타당성을 직접 검토하는 실습 자료로 활용될 수 있다. 또한 본 연구 결과는 치위생(학)과 교육과정에서 생성형 AI 활용 지침을 수립할 때, 단순 정답 확인 용도(암기형 문항)와 비판적 검토가 필수적인 용도(추론형, 법규 영역)를 구분하여 안내하는 근거로 활용될 수 있다. 한편 국가시험 외 일반 교과목에서도 LLM 기반 수업 운영 사례에서 역할 부여나 단계적 사고 유도와 같은 프롬프트 기법이 학습자의 AI 활용 능력 향상에 기여할 수 있음이 보고되고 있으며(23), 본 연구에서 확인된 프롬프트 전략별 반응 차이는 이러한 교육 맥락의 프롬프트 활용 원리와도 맥락을 같이한다.

본 연구는 다음과 같은 한계점을 갖는다. 첫째, 단일 모델(GPT-4o-2024-11-20), 단일 시점, 단일 Temperature(=0) 조건에서 수행된 결과로, 다른 모델이나 버전에서는 결과가 달라질 수 있다. 또한 본 연구는 GPT-4o를 대상으로 하였으나, 추론 특화 모델(예: OpenAI o1)에서는 추가적인 정답률 향상이 보고된 바 있어(24), 향후 연구에서는 이러한 최신 모델까지 포함한 비교가 필요하며, 본 연구에서 수치상 가장 높은 정답률을 보인 P5를 비교의 기준 프롬프트로 활용할 수 있을 것이다. 둘째, P5(통합 프롬프트)는 역할 부여, 근거 요청(CoT), Few-shot이라는 세 가지 요소를 동시에 적용한 결과로, 어떤 개별 요소 또는 요소 간 상호작용이 정답률 향상에 주된 영향을 미쳤는지는 본 연구의 설계로는 직접 분리하여 확인할 수 없다. 향후 요인설계(Factorial design)를 통한 후속 연구가 필요하다. 셋째, 본 연구는 한국어로 출제된 한국 치과위생사 국가시험을 대상으로 하였다는 점에서 결과의 언어적, 문화적 특수성이 있을 수 있다. 다만 이는 동시에, 영어권 중심으로 이루어져 온 기존 LLM-국가시험 평가 연구의 공백을 메우는 자료로서 의미를 가질 수 있다. 넷째, 그림 또는 사진 등 시각 자료가 포함된 문항은 분석에서 제외되었고 동일 문항에 대한 반복 측정(Test-retest)이 수행되지 않아, LLM의 멀티모달 성능 및 응답의 회차 간 변동성(Variability)은 본 연구의 결과에 반영되지 않았다. 다섯째, 본 연구에서 사용한 기출문항은 공개된 자료에 기반하므로, 2023년 문항을 Few-shot 예시로만 활용하고 평가 대상에서 제외하는 방식으로 실험 내 직접적인 문항 누출은 최소화하였으나, GPT-4o의 사전 학습 데이터에 해당 문항이 포함되어 있을 가능성 자체를 완전히 통제하기는 어려우며, 이 점은 결과 해석 시 고려해야 할 제한점이다. 마지막으로, 본 연구의 오류 패턴 분석은 반복적으로 관찰된 대표 사례 4개 유형의 기술적 제시에 초점을 두었으며, 전체 오답 문항에 대한 체계적 분류나 Thematic analysis는 수행하지 않았다. 향후 연구에서는 다중 LLM 간 프롬프트 전략 효과의 체계적 비교, 추론 특화 모델 및 국내 모델 포함, 멀티모달 문항 평가, 오류 유형의 체계적 Taxonomy 개발, 그리고 P5를 기준 프롬프트로 활용한 다모델 비교 연구가 필요하다.

5. 결론

본 연구는 2024–2025년 치과위생사 국가시험 기출문항 399개를 대상으로, GPT-4o에 대해 다섯 가지 프롬프트 전략의 정답률을 비교하였다. 그 결과, 모든 프롬프트 조건에서 전체 총점 기준 합격선을 상회하는 정답률을 확인하였으나, 5개 프롬프트 조건 간 전체 정답률 차이는 통계적으로 유의하지 않았다. 추론형 문항에서는 통합 프롬프트(P5)가 수치상 가장 높은 정답률을 보이는 경향이 있었으나, 통계적으로 확증되지는 않아 향후 검증이 필요하다.

이러한 결과는 치과위생사 및 치위생(학)과 학생이 생성형 AI를 활용할 때, 프롬프트 전략만으로 정답률이 크게 개선되기를 기대하기보다, AI가 제시하는 응답과 근거를 비판적으로 검토하는 것이 중요함을 시사한다. 오답 패턴 분석을 통해 확인된 한국 특화 지식 영역에서의 환각, 개념 혼동, 임상판단 오류는 생성형 AI 활용 시 비판적 검토 능력, 즉 AI 리터러시 함양 교육의 필요성을 보여준다.

Acknowledgements

본 연구는 자료 분석 과정에서 생성형 AI (Claude, Anthropic)를 보조적으로 활용하였으며, 최종 분석 결과 및 해석에 대한 책임은 저자에게 있다.

Author Contributions

Conceptualization: SK Kim. Data curation: SK Kim. Formal analysis: SK Kim. Investigation: SK Kim. Methodology: SK Kim. Writing-original draft: SK Kim. Writing-review & editin: SK Kim.

References

1. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023;11(6):887. 10.3390/healthcare11060887.
2. Xu X, Chen Y, Miao J. Opportunities, challenges, and future directions of large language models, including ChatGPT in medical education: a systematic scoping review. J Educ Eval Health Prof 2024;21:6. 10.3352/jeehp.2024.21.6.
3. Bae SM, Jeon HR, Kim GN, Kwak SH, Lee HJ. Performance of ChatGPT on the Korean national examination for dental hygienists. J Dent Hyg Sci 2024;24(1):62-70. 10.17135/jdhs.2024.24.1.62.
4. Song ES, Lee SP. Comparative analysis of the response accuracies of large language models in the Korean national dental hygienist examination across Korean and english questions. Int J Dent Hyg 2025;23(2):267-276. 10.1111/idh.12848.
5. Kim WJ, Kim BC, Yeom HG. Performance of large language models on the Korean dental licensing examination: a comparative study. Int Dent J 2025;75(1):176-184. 10.1016/j.identj.2024.09.002.
6. Song ES, Kim GH, Lee SP. Evaluation of GPT-4o and Gemini advanced on the Korean national dental licensing examination: accuracy, consistency, and question generation. J Dent Sci 2026;21(1):96-102. 10.1016/j.jds.2025.07.020.
7. Choi EM, Park SK. Comparative performance of GPT-4o and GPT-o1 on the Korean dental hygienist licensing examination. Journal of The Health Care and Life Science Sci 2025;13(1):1-8. 10.22961/JHCLS.2025.13.1.001.
8. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst 2020;33:1877-1901. 10.48550/arXiv.2005.14165.
9. Shanahan M, McDonell K, Reynolds L. Role play with large language models. Nature 2023;623(7987):493-498. 10.1038/s41586-023-06647-8.
10. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. Adv Neural Inf Process Syst 2022;35:24824-24837. 10.48550/arXiv.2201.11903.
11. Jang D, Yun TR, Lee CY, Kwon YK, Kim CE. GPT-4 can pass the Korean national licensing examination for Korean medicine doctors. PLOS Digit Health 2023;2(12):e0000416. 10.1371/journal.pdig.0000416.
12. Hsieh MY, Wang TL, Su PH, Chou MC. Impact of prompt engineering on the performance of ChatGPT variants across different question types in medical student examinations: cross-sectional study. JMIR Med Educ 2025;11:e78320. 10.2196/78320.
13. Kim B, Kang J, Kim MY, Ahn J. Prompt engineering for single-best-answer multiple-choice questions in licensing examinations: a narrative review with a case study involving the Korean medical licensing examination. J Educ Eval Health Prof 2025;22:34. 10.3352/jeehp.2025.22.34.
14. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198. 10.1371/journal.pdig.0000198.

15. Korea health personnel licensing examination institute. examination information by occupation: dental hygienist. [online] https://www.kuksiwon.or.kr. 2026.6.12.
16. Ministry of Health and Welfare. Enforcement Rule of the Act on the Persons Engaged in Medical Service, Article 9 (Determination of Successful Applicants, etc.), Paragraph 1. Korea Ministry of Government Legislation. [online] https://www.law.go.kr. 2026.6.12.
17. Miyazaki Y, Hata M, Omori H, Hirashima A, Nakagawa Y, Eto M, et al. Performance of ChatGPT-4o on the Japanese medical licensing examination: evaluation of accuracy in text-only and image-based questions. JMIR Med Educ 2024;10:e63129. 10.2196/63129.
18. Morishita M, Fukuda H, Yamaguchi S, Muraoka K, Nakamura T, Hayashi M, et al. An exploratory assessment of GPT-4o and GPT-4 performance on the Japanese national dental examination. Saudi Dent J 2024;36(12):1577-1581. 10.1016/j.sdentj.2024.11.006.
19. Maharjan J, Garikipati A, Singh NP, Cyrus L, Sharma M, Ciobanu M, et al. OpenMedLM: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. Sci Rep 2024;14(1):14156. 10.1038/s41598-024-64827-6.
20. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans Inf Syst 2025;43(2):1-55. 10.1145/3703155.
21. Asgari E, Montaña-Brown N, Dubois M, Khalil S, Balloch J, Au Yeung J, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. NPJ Digit Med 2025;8(1):274. 10.1038/s41746-025-01670-7.
22. Long D, Magerko B. What is AI literacy? competencies and design considerations. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020;1-16. 10.1145/3313831.3376727.
23. Naamati-Schneider L, Alt D. Developing AI literacy in medical and health professions education: exploring ChatGPT, ChatBots development, and simulation-based learning. Nurse Educ Pract 2025;88:104572. 10.1016/j.nepr.2025.104572.
24. Benito P, Isla-Jover M, González-Castro P, Fernández Esparcia PJ, Carpio M, Blay-Simón I, et al. GPT-4o and OpenAI o1 performance on the 2024 spanish competitive medical specialty access examination: cross-sectional quantitative evaluation study. JMIR Med Educ 2026;12:e75452. 10.2196/75452.