

기질 및 성격 기반 가상환자 생성 및 응답 비교 시스템 개발에 관한 연구

최영미¹ · 허수진² · 김성희^{3*}

¹동의대학교 인공지능학과 석사과정

²동의대학교 인공지능학과 박사과정

³동의대학교 인공지능학과 부교수

Developing a Temperament- and Character-Based System for Virtual Patient Generation and Response Comparison

Young-Mi Choi¹ · Sujin Hur² · Sung-Hee Kim^{3*}

¹Master's Student, Department of Artificial Intelligence, Dong-eui University, Busan 47340 Korea

²Ph.D. Student, Department of Artificial Intelligence, Dong-eui University, Busan 47340 Korea

³Associate Professor, Department of Artificial Intelligence, Dong-eui University, Busan 47340 Korea

[요약]

LLM 기반 가상환자는 임상 의사소통 훈련에 활용될 수 있으나, 페르소나 설계와 응답 검토에는 부담이 따른다. 본 연구는 기질 및 성격검사(TCI)를 설계 기준으로 활용해 페르소나 구성과 응답 비교를 지원하는 시스템을 제안하였다. 44개 페르소나와 10개 질문으로 세 차례 1,320개 응답을 생성하고, 동일 생성조건의 2·3회차 880개를 분석하였다. 반복 일관성은 SBERT와 Gemini 의미 판정으로, 구별성은 동일·상이 페르소나 간 비교로 평가하였다. 의미 유지율은 93.41%, 평균 SBERT 유사도는 0.8289, 완전 일치는 46.59%였다. 약물 관련 사실 충돌이 확인되었으며, 페르소나 구별성은 제한적이었다. 제안 시스템의 안정적인 활용을 위해 고정 임상정보와 행동 규칙의 강화가 필요하다.

[Abstract]

LLM-based virtual patients can support clinical communication training, but persona design and response review remain burdensome. This study proposes a Temperament and Character Inventory-based system for persona design, response generation, comparison, and visualization. Across three runs, 44 personas answered 10 questions, generating 1,320 responses; 880 responses from Runs 2 and 3 with identical generation conditions were analyzed. Consistency was assessed using SBERT similarity and Gemini-based semantic judgments, whereas persona distinctiveness was examined through within- and between-persona comparisons. Semantic preservation was 93.41%, mean SBERT similarity was 0.8289, and complete agreement was 46.59%. Medication-related factual conflicts occurred, and persona distinctiveness was limited. Reliable and safe use in educational settings requires stronger control of fixed clinical facts and persona-specific behavioral rules.

색인어 : 가상환자, 기질 및 성격검사, 성격 기반 페르소나, 대규모 언어 모델, 임상 의사소통 교육

Keyword : Virtual Patient, Temperament and Character Inventory, Personality-Based Persona, Large Language Model, Clinical Communication Education

<http://dx.doi.org/10.9728/dcs.2026.27.9.2573>



This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received 24 July 2026; **Revised** 14 August 2026

Accepted 08 September 2026

***Corresponding Author; Sung-Hee Kim**

Tel: +82-51-890-2366

E-mail: sh.kim@deu.ac.kr

I. 서 론

임상 교육에서 환자와의 의사소통 능력과 임상적 판단 능력은 환자의 상태와 요구를 파악하고 적절하게 대응하기 위해 요구되는 핵심 역량이다[1]. 이러한 역량은 임상 지식의 습득만으로 충분히 형성되기 어려우며, 실제 임상 상황을 모사한 반복적 연습과 피드백이 필요하다[2],[3]. 이에 따라 임상 교육에서는 동료 간 롤플레이팅(Peer role-playing)과 표준화 환자(Standardized patient)를 활용한 시뮬레이션 기반 교육이 활용되어 왔다[4],[5].

그러나 동료 간 롤플레이팅은 참여자의 역할 이해도와 몰입 수준에 따라 환자 표현과 상호작용 내용이 달라질 수 있으며, 표준화 환자는 역할자 모집과 훈련, 일정 조율 및 운영에 지속적인 인력과 비용이 요구된다[4],[6]. 이러한 운영적 제약은 제한된 교육 자원 안에서 다양한 환자 사례를 반복적으로 제공하는 데 부담으로 작용하며, 학습자가 서로 다른 성격과 반응 특성을 가진 환자를 충분히 경험하는 데 한계가 될 수 있다.

최근에는 이러한 한계를 보완하기 위해 LLM(Large Language Model)을 활용한 가상환자가 주목받고 있다[7],[8]. LLM 기반 가상환자는 학습자의 자유로운 질문에 응답할 수 있으며, 환자의 기본정보와 임상 상황, 응답 태도 등을 프롬프트에 반영하여 다양한 가상환자 페르소나를 구성할 수 있다[8],[9]. 또한 시간과 장소의 제약을 줄이고 반복적인 임상 면담 훈련을 지원할 수 있다는 점에서 기존 시뮬레이션 방식의 보완 수단으로 활용될 가능성이 제시되고 있다[7],[8].

그러나 교육 목적에 맞는 가상환자 페르소나를 구성하기 위해서는 환자의 기본정보와 임상 사례뿐만 아니라 성격, 정보 제공 범위, 대화 태도 및 응답 방식 등 다양한 요소를 프롬프트에 반영해야 한다[7]-[9]. 이러한 요소의 구성과 표현 방식은 연구 목적과 시스템에 따라 달라질 수 있으며[7],[8],[10], 설정한 특성이 생성 응답에 적절히 반영되었는지 확인하기 위해 반복적인 응답 생성과 검토, 프롬프트 수정이 요구될 수 있다[11]. 특히 동일한 조건에서 생성된 응답이라도 회차에 따라 표현이나 세부 내용이 달라질 수 있으므로, 설정된 페르소나 특성이 반복 생성 과정에서도 유지되는지 확인할 수 있는 평가 절차가 필요하다. 이러한 과정은 다수의 가상환자를 설계하고 비교하려는 교수자에게 추가적인 부담으로 작용할 수 있다.

이에 본 연구에서는 가상환자 설계 과정에서 나타나는 프롬프트 구성의 모호성, 교수자의 프롬프트 작성 역량에 대한 의존, 반복적인 응답 생성 및 검토 과정의 부담을 완화하고자 한다. 이를 위해 Cloninger의 기질 및 성격 이론(TCI: Temperament and Character Inventory)[12]을 페르소나 설계 기준으로 활용하여, 페르소나 구성과 응답 생성·분석·비교를 통합적으로 지원하는 시스템을 제안하고 프로토타입으

로 구현하였다. 또한 동일한 페르소나와 질문에 대해 반복 생성된 응답을 비교하여, 설정된 특성과 핵심 내용이 회차 간 일관되게 유지되는지를 검토하였다.

이를 바탕으로 다음과 같은 연구질문을 설정하였다. ‘RQ1. 동일한 TCI 설계조건인 페르소나에서 반복 생성된 응답은 어느 정도 의미적 일관성을 보이는가?’, ‘RQ2. 서로 다른 TCI 설계조건인 페르소나에서 생성된 응답은 어느 정도 구별 가능한가?’ 이를 통해 제안 시스템의 적용 가능성과 향후 가상 환자 설계 및 평가 절차의 확장 방향을 제시한다.

II. 관련 연구

2-1 임상 시뮬레이션의 기존 방식과 한계

임상 교육에서 환자와의 의사소통 능력과 임상적 판단 능력은 환자의 상태나 요구를 파악하고 적절히 대응하기 위해 갖추어야 할 핵심 역량 중 하나이다[1]. 이러한 역량은 이론적 지식만으로는 충분히 형성되기 어려우므로, 실제 상황에서 지식과 의사소통 기술, 임상적 판단을 통합적으로 적용하는 경험이 필요하다[2],[3].

그러나 임상 교육 과정에서 학습자가 실제 환자를 대상으로 충분한 면담과 의사결정 경험을 반복적으로 경험하는 데에는 환자 안전, 제한된 교육 기회 및 임상 환경의 불확실성 같은 제약이 따른다[13],[14]. 이에 따라 실제 임상 상황을 안전하게 모사하면서 반복적인 연습과 피드백을 제공할 수 있는 시뮬레이션 기반 교육이 임상 역량 훈련의 주요 방법으로 활용되어 왔다[2],[3].

임상 시뮬레이션의 대표적인 방식에는 학습자가 환자와 의료인의 역할을 번갈아 수행하는 동료 간 롤플레이팅과, 훈련된 역할자가 임상 사례를 일정하게 재현하는 표준화 환자가 있다[4],[5],[15],[16]. 두 방식은 실제 임상 상황과 유사한 환경에서 의사소통과 임상 수행을 반복적으로 훈련할 수 있으나[5],[17], 동료 간 롤플레이팅은 참여자의 역할 이해도와 몰입 수준에 따라 상호작용의 일관성이 달라질 수 있고[4],[6], 표준화 환자는 사례 개발, 역할자 모집·훈련, 일정 조율 및 수행 관리에 지속적인 인력과 비용이 요구된다[6],[18]. 이러한 제약은 다양한 환자 사례를 반복적이고 대규모로 제공하는 데 부담으로 작용한다.

이러한 한계를 보완하기 위해 최근에는 시간과 장소의 제약을 줄이고 반복적인 임상 면담 훈련을 지원할 수 있는 LLM 기반 가상환자가 주목받고 있다[7],[8],[19]. LLM 기반 가상환자는 자연어 생성 능력을 바탕으로 학습자의 자유로운 질문에 즉시 응답할 수 있으며, 환자의 기본정보와 임상 상황, 응답 태도와 같은 조건을 프롬프트에 반영하여 다양한 면담 사례를 구성할 수 있다[7],[8].

2-2 LLM기반 가상환자 설계의 모호성과 검토 부담

LLM 기반 가상환자를 수업에 활용하기 위해서는 교수자가 학습 목표와 임상 상황에 따라 환자의 기본정보, 증상과 병력, 대화 태도 및 응답 방식 등을 프롬프트에 구체적으로 제시해야 한다. 그러나 기존의 LLM 기반 가상환자 구현 연구에서는 프롬프트에 포함하는 요소와 구성 방식이 연구 목적과 시스템에 따라 서로 다르게 나타난다[7],[8],[10],[20]. 또한 동일한 정보를 포함하더라도 지시문의 표현, 정보 제시 순서 및 프롬프트 형식에 따라 생성 결과가 달라질 수 있다[21],[22]. 이처럼 프롬프트에 포함되는 요소와 작성 방식이 다양하고, 세부적인 표현에 따라 생성 결과가 달라질 수 있어 교수자가 수업 목적에 맞는 환자 특성과 응답 방식을 어떤 형태로 구체화해야 하는지 판단하기 어렵다.

이러한 설계의 어려움으로 인해 교수자는 생성된 응답이 의도한 환자 특성과 수업 목적에 부합하는지 확인해야 한다. 이를 위해 여러 임상 질문에 대한 응답을 생성하고, 설정과 다른 반응이 나타나는 경우에는 프롬프트를 수정한 뒤 다시 생성·검토하는 과정을 반복해야 한다[7],[11]. 특히 페르소나와 질문의 수가 증가할수록 검토해야 할 응답도 함께 늘어나며, 개별 응답을 확인하는 방식만으로는 페르소나별 특성과 차이를 파악하기 어려워진다. 따라서 다수의 생성 응답을 체계적으로 정리하고 페르소나 간 반응 차이를 비교할 수 있는 지원이 필요하다.

2-3 페르소나 설계를 위한 TCI 기반 접근

가상환자의 성격을 구성할 때 ‘예민한 환자’나 ‘소극적인 환자’와 같은 일상적 표현에만 의존하면, 동일한 표현이라도 교수자에 따라 다르게 해석될 수 있으며 이를 구체적인 행동이나 응답 방식으로 변환하는 과정에서도 차이가 발생할 수 있다. 따라서 가상환자에 적용할 성격을 공통된 개념과 차원에 따라 명시하기 위해서는, 개인의 반응 경향을 체계적으로 설명할 수 있는 성격 이론에 기반한 접근이 필요하다[23],[24].

이에 본 연구에서는 Cloninger의 TCI 이론[12]을 활용하였다. TCI는 개인의 성격을 기질과 성격의 두 영역으로 구분한다. 기질은 자극추구(NS; Novelty Seeking), 위험회피(HA; Harm Avoidance), 사회적 민감성(RD; Reward Dependence), 인내력(P; Persistence)의 네 차원으로 구성되며, 성격은 자율성(SD; Self-directedness), 연대감(CO; Cooperativeness), 자기초월(ST; Self-transcendence)의 세 차원으로 구성된다. TCI는 개인의 반응 경향을 여러 차원의 조합으로 설명할 수 있어 다양한 페르소나 조건을 체계적으로 구성하는 이론적 기준으로 활용할 수 있다. 또한 위험회피, 자극추구, 인내력 및 자율성 등 여러 TCI 차원과 우울 간의 관련성이 보고되어 왔으며[25]–[28], 본 연구에서는 초기 임상 사례로 우울증을 선정한 점을 고려하여 TCI를 페르

소나 설계를 위한 초기 설계 틀로 활용하였다. 다만 TCI는 심리측정학적 성격을 직접 측정하거나 구현하기 위한 도구가 아니라, 서로 다른 페르소나 조건을 구성하기 위한 설계 기준으로 사용하였다.

III. 시스템 설계 및 구현

본 연구에서는 TCI 기반 가상환자 설계 시스템을 제안하고 프로토타입으로 구현하였다. 제안 시스템은 일상적 표현에 의존하던 환자의 성격을 TCI에 따라 조합하고, 서로 다른 페르소나에서 생성된 응답을 비교·검토하는 과정을 지원하는 데 목적이 있다.

그림 1과 같이 제안 시스템은 페르소나 설계(Persona Design), 응답 생성(Response Generation), 응답 분석 및 비교(Response Analysis and Comparison), 비교 시각화(Comparative Visualization)의 네 단계로 구성된다. 먼저 페르소나 설계 단계에서는 사용자가 환자의 기본정보를 입력하면, 시스템이 사전에 정의된 TCI 조합을 각각 적용하여 서로 다른 페르소나를 구성한다. 응답 생성 단계에서는 각 페르소나에 동일한 임상 질문을 적용하여 가상환자의 응답을 생성한다. 응답 분석 및 비교 단계에서는 생성된 응답의 정서적·행동적·언어적 양상을 분석하고 페르소나 간 차이를 비교한다. 마지막으로 비교 시각화 단계에서는 분석 결과를 다양한 시각화 방식으로 제공하여 페르소나별 응답 양상을 확인할 수 있도록 한다.

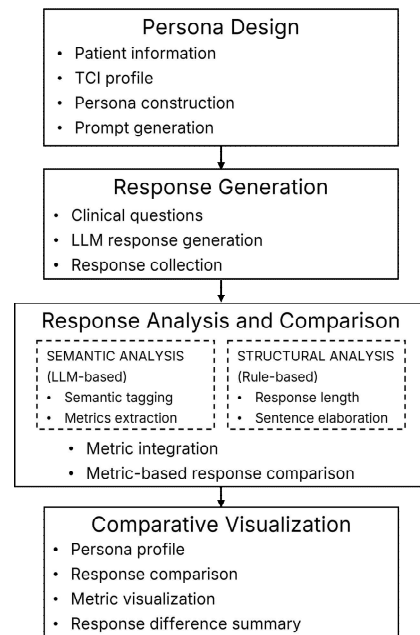


그림 1. 제안 시스템의 아키텍처

Fig. 1. Architecture of the proposed system

3-1 페르소나 설계

본 연구에서는 제안 시스템의 프로토타입을 구현하기 위한 초기 임상 사례로 우울증을 선정하였다. 본 연구에서는 실제 환자의 의무기록이나 개인정보를 사용하지 않았으며, 연구자가 가상으로 설정한 환자 기본정보와 임상 조건을 사용하였다. 페르소나 설계에 적용할 TCI 수준의 구분에는 최지영[28]의 우울장애 환자 대상 연구에서 사용한 세 수준 구분 방식을 참고하였다. 최지영[28]은 우울장애 환자의 TCI 기질 특성을 분석하는 과정에서 백분위 점수의 30백분위와 70백분위를 기준으로 기질 수준을 낮음, 중간, 높음으로 구분하였다. 본 연구에서는 이와 같은 세 수준의 구분 방식을 참고하되, 해당 백분위 기준을 실제 TCI 검사점수의 절단점으로 적용하지 않았다. 대신 가상환자의 기질 및 성격 특성을 프롬프트에서 상대적으로 구분하기 위한 설계 수준값으로 상(High), 중(Middle), 하(Low)를 사용하였다. 따라서 본 연구에서 사용한 상·중·하는 실제 대상자의 TCI 검사점수에 따른 심리적 정학적 분류를 의미하지 않는다.

표 1. 본 연구에서 구성한 TCI 기반 페르소나 조합

Table 1. TCI-based persona combinations constructed in this study

Temperament				
NS	HA	RD	P	Descriptive Persona Label
Middle	High	Low	Low	Socially Withdrawn and Fearful
High	High	Low	Low	Impulsive and Explosive
High	High	High	High	Hypersensitive
Middle	Middle	Middle	Middle	Adaptive and Flexible
Character				
SD	CO	ST	Descriptive Persona Label	
Low	Low	Middle	Emotionally Immature	
Low	High	Middle	Submissive	
Low	Middle	Low	Imitative	
Low	Middle	High	Disorganized Thinking	
Low	Low	Low	Dysphoric	
Low	Low	High	Disorganized	
Low	High	Low	Dependent	
Low	High	High	Emotionally Reactive	

표 1에 제시한 기질 및 성격 차원의 수준을 조합하여 기질 단독 조합 4개와 성격 단독 조합 8개를 구성하였다. 제시한 페르소나 명칭은 TCI에서 공식적으로 정의한 성격유형이 아니라, 각 TCI 차원의 조합을 설명하기 위해 연구자가 부여한 기술적 명칭이다. 이들 조합은 TCI에서 정의된 모든 가능한 수준 조합을 구성한 것이 아니라, 서로 다른 기질 및 성격 조건에 따른 응답 양상을 비교하기 위해 연구자가 구성한 설계 조건이다. 이후 두 조합군을 교차 결합하여 32개의 복합 조합을 추가하였으며, 최종적으로 총 44개의 페르소나를 구성하

였다.

구성된 페르소나 조합은 사용자가 입력한 이름, 나이, 성별 및 주소증 등의 환자 기본정보와 결합하여 가상환자 응답 생성용 프롬프트로 변환하였다. 기질 단독 조합의 프롬프트에는 기질 영역만 포함하고 성격 영역은 제외하였으며, 성격 단독 조합에서도 성격 영역만 포함하였다.

표 2. TCI 기반 가상환자 생성을 위한 프롬프트 템플릿

Table 2. Prompt template for TCI-based virtual patient generation

프롬프트
당신은 다음 조건을 가진 가상환자입니다. 이 역할을 완전히 수행해주세요.
[환자 정보] - 이름: {이름} - 나이: {나이} - 성별: {성별} - 주소증: {주소증}
[TCI 기반 성격 특성] - 기질(Temperament) · 자극추구: {수준} · 위험회피: {수준} · 사회적인감성: {수준} · 인내력: {수준} - 성격(Character) · 자율성: {수준} · 연대감: {수준} · 자기초월: {수준}
[대화 규칙] 1. 위의 기질과 성격 조건을 심리적 특성으로 가정하고, 이에 맞는 말투, 감정 표현 및 사고방식이 자연스럽게 드러나도록 대답하세요. 단, 해당 조건을 응답에서 직접 설명하지 마세요. 2. 모든 답변은 실제 상담을 받는 환자의 1인칭 시점으로 작성하세요. 3. 감정이 강한 상태에서는 짧고 직접적인 표현을 사용하고, 감정이 상대적으로 약한 상태에서는 상황을 설명하는 방식으로 답하세요. 4. 망설이거나 말을 흐리는 표현은 상황에 맞을 때만 제한적으로 사용하고, 같은 형태를 반복하지 마세요. 5. 매번 동일한 방식으로 응답을 시작하지 말고, 질문과 상황에 맞는 방식으로 답변을 시작하세요. 6. 의학적·전문적 표현보다 일상적인 표현을 사용하세요. 7. 실제 진료 상황에서 환자가 의료진에게 말하는 것처럼 솔직하고 차분한 말투를 유지하세요.
이제 위 환자의 입장에서 다음 질문에 대답하세요.
[질문] {임상 질문}

*The prompt was written in Korean.

표 2는 가상환자 응답 생성에 사용한 공통 프롬프트 템플릿을 나타낸다. 프롬프트에는 환자 기본정보와 TCI 조건 외에도 가상환자가 환자의 관점을 유지하고 설정된 특성을 실제 대화에 반영할 수 있도록 응답 관점과 표현 방식에 관한 대화 규칙을 포함하였다. 또한 질문에 대해 환자의 입장에서 직접 응답하도록 역할을 지정하고, 설정과 무관한 설명이나

제3자 관점의 답변을 생성하지 않도록 하였다. 본 연구에서는 각 TCI 차원을 특정 행동이나 언어 표현에 일대일로 대응시키는 별도의 규칙을 정의하지 않았으며, 제시된 TCI 수준과 공통 대화 규칙을 바탕으로 생성 모델이 해당 조건에 따라 응답하도록 하였다. 이러한 공통 규칙은 모든 페르소나에 동일하게 적용하였다. 따라서 본 연구에서 직접 비교한 것은 연구자가 설정한 TCI 차원의 수준값을 프롬프트 조건으로 적용했을 때 생성되는 응답의 차이이며, TCI의 심리학적 구성개념이 LLM의 언어·행동 특성으로 타당하게 구현되었는지를 검증한 것은 아니다. 또한 공통 대화 규칙의 영향과 TCI 조건의 영향을 독립적으로 분리하지 않았으므로, 관찰된 응답 차이를 개별 TCI 차원의 독립적인 효과로 해석하는 데에는 제한이 있다.

3-2 가상환자 응답 생성

페르소나별 응답을 비교하기 위해 기초 임상 문진 항목과 우울증 한의표준임상진료지침의 주요 동반 증상 평가 항목을 참고하여 10개의 질문을 구성하였다[29],[30]. 표 3과 같이 질문은 주소증, 만성질환과 건강 문제, 약물 복용, 자살사고, 정신과 치료 경험, 식욕 변화, 삶의 이유, 절망감, 불안과 우울, 충동적 행동에 관한 내용을 포함하였다. 표 3의 질문은 참고문헌 [29], [30]의 문항을 직접 인용하거나 번역한 것이 아니라, 관련 임상 문진 및 평가 내용을 바탕으로 본 연구의 가상환자 응답 비교 목적에 맞게 재구성하였다.

가상환자 응답 생성에는 OpenAI의 gpt-4.1-mini를 사용하였다. 생성 당시 세부 모델 snapshot은 별도로 기록하지 않아 API 요청에 지정한 모델명 수준에서 기술하였으며, 이에 따라 동일 모델명을 사용하더라도 완전히 동일한 결과의 재현에는 제한이 있다. 응답 생성 시 최대 토큰 수는

표 3. 페르소나 응답 생성을 위한 임상 질문

Table 3. Clinical questions for persona responses

No	Content
1	What brings you in today?
2	Do you have any pre-existing medical conditions or health problems?
3	Are you currently taking any medications?
4	During the past month, have you had any thoughts of suicide?
5	Have you ever received psychiatric treatment?
6	During the past month, have you experienced any changes in your appetite?
7	What are your reasons for living?
8	During the past month, have you felt hopeless, as though your situation would not improve?
9	During the past month, have you felt anxious or depressed?
10	Do you often engage in impulsive behavior?

max_tokens=500으로 설정하였고, temperature, top_p 및 seed는 별도로 지정하지 않아 API 기본 설정을 사용하였다.

44개 페르소나에 10개의 질문을 각각 적용하여 회차당 440개의 응답을 생성하였고, 이를 세 차례 수행하여 총 1,320개의 응답을 수집하였다. 본 연구의 반복 생성은 탐색적으로 세 차례 수행하였으며, 반복횟수에 대한 별도의 통계적 산정은 수행하지 않았다. 회차별 생성은 OpenAI Batch API를 통해 별도의 작업으로 수행하였으며, 이전 회차의 생성 응답은 이후 회차의 입력에 포함하지 않았다. 응답 생성은 2026년 7월 16일부터 17일까지 수행하였다. 사후 데이터 무결성 점검 결과, 1회차에서 환자 기본정보가 API 요청에 정상적으로 반영되지 않은 것이 확인되어, 분석에는 동일한 생성조건이 확인된 2회차와 3회차의 응답을 사용하였다.

그림 2는 환자 기본정보를 입력하고 가상환자 응답을 생성·검토하는 인터페이스를 나타낸다. 그림 2a에서는 환자의

*The interface is presented in Korean because it was designed for Korean-speaking users.

그림 2. 가상환자 정보 입력, 페르소나 선택 및 응답 프롬프트 확인 구현 인터페이스

Fig. 2. Interface for patient input, persona selection, and response prompt review

이름, 나이, 성별 및 주소증을 입력하고 생성 세션을 관리하며, 입력된 환자 기본정보에 사전 구성된 TCI 기반 페르소나 조합이 적용된다. 공통 임상 질문을 선택하면 그림 2b와 같이 동일한 질문에 대한 44개 페르소나의 응답과 기질 및 성격 설정이 카드 형태로 제시된다. 사용자는 응답을 검토하고 특정 페르소나를 선택하여, 그림 2c에서 선택한 가상환자에 적용된 환자 기본정보, TCI 설정 및 대화 규칙을 확인할 수 있다. 이를 통해 생성된 응답과 적용된 페르소나 조건을 함께 검토하고, 이후 분석 및 비교에 사용할 응답을 선택할 수 있도록 하였다.

3-3 가상환자 응답 분석

본 연구에서는 페르소나별 생성 응답의 표현 양상과 텍스트 구조를 탐색적으로 비교하기 위해 점수형 지표 8개와 텍스트 구조 지표 5개로 총 13개의 분석 지표를 구성하였다. 점수형 지표는 생성 응답에서 나타나는 정서적 표현, 반응의 주체성 및 대인 지향적 표현, 발화의 기능적 유형, 해결 관련 표현, 응답의 길이와 구조적 복잡성을 서로 다른 측면에서 비교·시각화하기 위해 선정하였다. 이들 지표는 생성 텍스트의 상대적 표현 차이를 확인하기 위해 연구자가 구성한 탐색적·기술적 응답 특성 지표로, TCI 차원이나 기타 심리학적 구성개념을 측정하는 심리적도를 의미하지 않는다. 각 지표의 가중치와 정규화 기준 역시 선행연구에서 검증된 심리측정학적 절단값이 아니라 생성 응답 간 상대적 차이를 비교하기 위해 연구자가 설정한 휴리스틱 기준이며, 사용한 키워드와 가중치 및 산출 기준은 표 4와 다음 절에 제시하였다.

점수형 지표에 사용한 키워드는 감정 강도의 고강도(행복, 기쁨, 만족, 희망, 감사, 사랑, 즐거움, 웃음, 슬픔, 화남, 우울, 불안, 짜증, 절망, 걱정, 스트레스), 중간(평온, 차분, 보통, 일반적, 평범), 낮음(무관심, 중립, 객관적, 사실적, 분석적), 정서 방향의 긍정(행복, 기쁨, 만족, 희망, 감사, 사랑, 즐거움, 웃음, 좋다, 훌륭하다)과 부정(슬픔, 화남, 우울, 불안, 짜증, 절망, 걱정, 스트레스, 나쁘다, 힘들다)으로 구성하였다. 주체성 관련 표현(Agency-related Expression)은 능동(할 것이다, 하려고, 노력, 의지, 목표, 계획, 결심, 직접, 스스로), 수동(될 것 같다, 아마도, 어쩌면, 그냥, 그대로, 기다린다), 의존(도움, 의존, 상담, 조언, 가르침, 지시, 명령) 표현으로 구분하였다. 대인 지향 표현(Social Orientation)은 타인·관계 지향(우리, 함께, 사회, 공동체, 협력, 소통, 이해, 공감, 친구, 가족)과 자기·개인 지향(나, 개인, 혼자, 자신, 내면, 사색, 고민, 생각) 표현으로 구분하였다. 표현 유형은 위로·격려(괜찮아, 힘내, 잘될 거야, 걱정마, 위로, 안심, 희망, 기대), 정보 제공(정보, 사실, 데이터, 통계, 연구, 분석, 결과, 증거), 지시(해야, 하지마, 필요해, 중요해, 당연해, 무조건), 해결 관련 표현은 직접적 표현(해결, 방법, 대안, 시도, 실행, 적응, 극복, 제안, 권장, 계획)과 간접적 표현(생각해보자, 고민해보자, 시간이 필요하다, 차근차근)으로 구성하였다.

표 4. 페르소나별 응답 분석 지표 및 산출 방식

Table 4. Persona response metrics and calculation methods

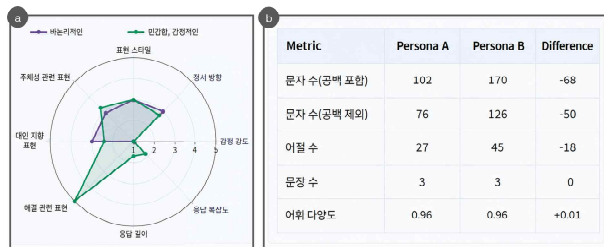
Analysis Type	Metric	Calculation and Value Type
Score-Based Metrics	Emotional Intensity, Valence, Agency-related Expression, Social Orientation	Predefined weighted keyword rules: Intensity (H/M/L=.4/.2/.3), Valence ($\pm$.5), Agency-related Expression (Active/Passive/Dependent=.6/.4/.5), Social Orientation (Interpersonal/Self-oriented=.5/.5); values limited to 1-5
	Expression Type, Solution-related Expression	Expression Type: Comfort/Information/Command=1/3/5; Solution-related Expression: Direct/Indirect/None=5/3/1
	Response Length, Complexity	Length=character count/100; Complexity=mean period-delimited segment length/30; values limited to 1-5
Text-Structure Metrics	Character Count (with spaces/without spaces), Word Count, Sentence Count	Directly calculated from the response text
	Lexical Diversity	Unique token count/total token count; 0-1

각 키워드는 응답 원문의 문자열 출현 횟수를 기준으로 계산하였다. 감정 강도는 고·중·저강도 키워드에 각각 0.4, 0.2, 0.3의 가중치를 적용하였으며, 정서 방향은 3점을 기준으로 긍정·부정 키워드의 가중 점수 차이를 반영하였다. 주체성 관련 표현과 대인 지향 표현도 3점을 기준으로 각 표현군의 가중 점수 차이를 반영하였다. 표현 유형은 위로·격려, 정보 제공 및 지시 표현의 가중 점수를 비교하여 우세한 유형에 따라 1점, 3점 또는 5점으로 구분하였으며, 해결 관련 표현은 직접적 표현이 있는 경우 5점, 간접적 표현만 있는 경우 3점, 관련 표현이 없는 경우 1점으로 구분하였다. 연속형 점수는 최종적으로 1점에서 5점 범위로 제한하였다. 이러한 키워드 기반 산출은 부정 표현, 문맥 및 다의어를 별도로 구분하지 않으므로, 각 지표는 실제 성격 특성이나 발화 의도를 직접 측정하는 값이 아니라 생성 응답의 상대적 언어표현을 비교하기 위한 기술적 값으로 해석하였다.

응답 길이는 공백을 포함한 전체 문자열 길이를 100으로 나누는 값으로 산출하였으며, 복잡도는 마침표를 기준으로 분할한 응답 단어의 평균 문자 길이를 30으로 나누어 산출한 후 각각 1점에서 5점 범위로 제한하였다. 텍스트 구조 지표는 공백 포함 글자 수, 공백 제외 글자 수, 단어 수, 문장 수 및 어휘 다양성으로 구성하였다. 공백 포함 및 제외 글자 수는 각각 원문 문자열과 공백을 제거한 문자열의 길이로 산출하였으며, 단어 수는 공백을 기준으로 분리된 단어의 수로 계산하였다. 문장 수는 마침표, 느낌표 및 물음표를 기준으로 분할된 문장 단어의 수로 산출하였고, 어휘 다양성은 전체 단어 수 대비 고유 단어 수의 비율로 계산하였다.

3-4 비교 시각화 및 인터페이스

앞서 생성·분석한 가상환자 응답을 대상으로, 본 시스템은 두 페르소나의 응답 특성을 직접 비교할 수 있도록 하였다. 사용자는 비교할 질문과 두 페르소나를 직접 지정하거나, 분석 결과를 기준으로 시스템이 비교 대상을 자동 선정하도록 설정할 수 있다. 자동 선정은 8개 점수형 지표에서 두 페르소나 간 절대 차이의 평균이 가장 큰 쌍을 선택하거나, 사용자가 8개 점수형 지표 중 하나를 선택하면 해당 지표에서 최고값과 최저값을 보이는 페르소나를 선택하는 방식으로 이루어진다.



*The text in the interface is presented in Korean

그림 3. 가상환자 응답 특성 비교 시각화 인터페이스

Fig. 3. Visualization interface for comparing virtual patient response characteristics

그림 3은 선정된 두 페르소나의 응답과 분석 결과를 비교하는 인터페이스를 나타낸다. 비교 화면에서는 각 페르소나의 환자 기본정보와 TCI 설정, 동일한 질문에 대한 응답 원문을 함께 제시하였다. 이를 통해 사용자는 서로 다른 TCI 조건이 적용된 응답의 내용과 표현 차이를 직접 확인할 수 있다.

분석 결과는 8개의 점수형 지표와 5개의 텍스트 구조 지표로 구분하여 제시하였다. 그림 3a의 레이더 차트는 감정 강도, 정서 방향, 표현 유형, 주체성 관련 표현, 대인 지향 표현, 해결 관련 표현, 응답 길이 및 복잡도의 값을 시각화한다. 레이더 차트에 제시된 값은 심리검사 결과나 페르소나의 성격 수준을 의미하지 않으며, 본 연구에서 정의한 탐색적 기술지표를 이용하여 생성 응답의 상대적 표현 차이를 시각화한 것이다. 그림 3b에서는 공백 포함 글자 수, 공백 제외 글자 수, 단어 수, 문장 수 및 어휘 다양성을 표 형태로 제시하고, 각 페르소나의 값과 차이를 함께 표시하였다. 이를 통해 사용자는 응답의 의미적·행동적 특성과 텍스트 구조를 함께 확인하면서 페르소나 간 차이를 비교·검토할 수 있다.

IV. 결 과

4-1 분석 자료 및 방법

생성 응답의 반복 일관성과 페르소나 간 구별 정도를 확인하기 위해 세 차례 생성한 총 1,320개의 응답 중 동일한 생성

조건이 확인된 2회차와 3회차의 880개 응답을 분석하였다. 반복 의미 일관성은 동일한 페르소나와 질문에서 2회차와 3회차에 생성된 응답을 대응시켜 총 440쌍을 구성하였다. 응답 간 의미 유사도는 sentence-transformers/paraphrase-multilingual-mpnet-base-v2를 이용한 SBERT 유사도로 산출하였다. SBERT 코사인 유사도는 두 응답 텍스트 간 의미적 유사도를 정량화하기 위한 지표로 사용하였으며, 임상적 타당성이나 TCI 구성개념의 동일성을 직접 측정하는 지표로 해석하지 않았다. 또한 임베딩 유사도만으로 확인하기 어려운 핵심 의미와 사실정보의 충돌을 평가하기 위해 2026년 7월 24일(KST)에 gemini-3.5-flash-lite를 사용하여 Gemini 기반 의미 관정을 수행하였으며, 세부 모델 snapshot은 별도로 기록되지 않았다. 평가는 Google Generative Language API의 v1beta generateContent 엔드포인트를 비동기 REST 방식으로 호출하여 수행하였고, 응답은 사전에 정의한 JSON schema에 따라 구조화하여 수집하였다. temperature, top_p, top_k 및 seed는 API 요청에서 별도로 지정하지 않았다. Gemini의 원 판정 범주는 일치, 부분 일치, 모순 및 판단 불가로 구성되었으며, 분석 결과에서는 ‘모순’을 ‘불일치’로 정규화하여 제시하였다. 일치는 핵심 사실과 답변 방향이 동일한 경우, 부분 일치는 두 응답이 양립할 수 있으나 세부 정보, 강도 또는 빈도 등에 차이가 있는 경우, 모순은 동일한 시간 범위와 사실에 대해 두 응답이 동시에 참일 수 없는 경우로 정의하였다. 판단 불가는 한쪽 또는 양쪽 응답이 질문에 답하지 않거나 비교가 어려운 경우로 하였다. 각 질문의 핵심 판단 요소를 반영한 세부 판정 기준을 적용하였으며, 평가에는 질문, 두 응답, 생성 회차, 페르소나 식별자와 해당 판정 기준을 제공하였다. SBERT 유사도와 연구자 판정은 제공하지 않았다. 두 응답은 항상 앞선 생성 회차를 응답 A로 제시하였고 별도의 순서 교환은 수행하지 않았으며, 각 응답쌍은 1회 판정하였다. Q10은 실제 문항에 맞는 충동적 행동 기준으로 수정하여 전체 132개 응답쌍을 재평가하였으며, 이 중 2·3회차의 44개 응답쌍을 주 분석에 반영하였다. 전체 평가 프롬프트, 질문별 판정 기준 및 구조화 출력 형식은 재현성을 위해 GitHub 저장소 [https://github.com/Choiyoungmi0122/PAPER_BATCH_VIS]에 제시하였다.

Gemini와 연구자 판정 간 일치 정도를 확인하기 위해 전체 Gemini 평가 결과에서 질문별 15쌍씩 층화 무작위 추출한 총 150쌍을 임상전문가가 아닌 석사과정 연구자인 제1저자 1인이 Gemini 판정과 SBERT 유사도를 확인하지 않은 상태에서 독립적으로 검토하였다. 이 중 Q10에 해당하는 인간 검토 표본 15쌍도 수정된 판정 기준에 따라 다시 검토하였으며, 인간 평가자가 1인이므로 평가자 간 일치도는 산출하지 않았다. 일치, 부분 일치 및 불일치를 순서형 범주로 간주하고 판정 차이의 정도를 반영하기 위해 quadratic weighted Cohen's κ 를 산출하였다. Gemini 판정 범주와 SBERT 유사도의 관계는 판정 범주와 질문을 설명변수로 포함한 선형회귀모형으로 분석하였으며, 동일 페르소나에서 생성된 응답의

군집성을 고려하여 페르소나 단위의 군집-강건 표준오차(cluster-robust standard error)를 적용하였다.

페르소나 구별성은 2회차와 3회차에서 생성된 동일 페르소나의 응답 440쌍과 서로 다른 페르소나의 응답 18,920쌍에 대한 SBERT 유사도를 비교하고, Top-1, Top-3, Top-5 정확도와 평균 역순위(MRR)를 산출하여 평가하였다. 서로 다른 페르소나 간 비교는 응답쌍의 의존성을 고려하여 기술적으로 비교하였다. 추가로 7개 TCI 차원이 모두 정의된 32개 복합 페르소나의 2·3회차 응답 640개를 대상으로 SBERT 임베딩 기반 선형 분류를 수행하였으며, 페르소나 단위 그룹 교차검증과 permutation baseline을 적용하고 다중검정에는 Holm 보정을 사용하였다.

4-2 반복 생성 응답의 의미 일관성

동일한 생성조건이 확인된 2회차와 3회차의 응답 440쌍을 비교한 결과는 표 5와 같다. 일치 46.59%, 부분 일치 46.82%, 불일치 6.59%로 나타났으며, 의미 유지율은 93.41%였다. 평균 SBERT 유사도는 0.8289 ± 0.0647 이었다. 반복 생성에서 핵심 의미는 대체로 유지되었으나 세부 정보까지 동일하게 재현되지는 않았다.

표 5. 반복 생성 응답의 전체 의미 일관성

Table 5. Overall semantic consistency of repeated responses

Evaluation type	Result
Gemini judgment	Agreement 46.59%; Partial Agreement 46.82%; Disagreement 6.59%
Semantic preservation	93.41%
SBERT similarity	0.8289 ± 0.0647

Gemini 판정 범주와 SBERT 유사도의 관계를 분석한 결과, 판정 범주의 효과는 통계적으로 유의하였다($F(2,43)=50.404$, $p<.001$). 일치를 기준으로 부분 일치의 SBERT 유사도는 0.0357, 불일치는 0.0876 낮게 나타났다. 질문별로는 전체 불일치 29건 중 Q3에서 19건, Q10에서 6건이 확인되었다.

Gemini 판정 절차의 예비적 일치 정도를 확인하기 위해 별도로 검토한 150쌍에서 연구자와 Gemini의 판정은 98쌍(65.3%)에서 일치하였으며, 이차 가중 Cohen's κ 는 0.4942였다. 다만 인간 검토가 임상전문가가 아닌 단일 연구자에 의해 수행되었다는 점에서, 이는 Gemini 판정의 임상적 타당성 보다는 본 연구에서 정의한 판정 기준에 대한 예비적 일치 정도로 해석하였다.

4-3 페르소나 간 응답 유사도 비교

동일한 생성조건이 확인된 2회차와 3회차 응답을 비교한 결과, 동일 페르소나 응답 440쌍의 평균 SBERT 유사도는 0.8289, 서로 다른 페르소나 응답 18,920쌍의 평균은

0.8189로 나타났으며, 두 조건 간 차이는 0.0100이었다. 그림 4에서도 동일 페르소나 응답의 유사도 분포가 상대적으로 높은 방향에 위치했으나, 두 분포는 상당 부분 중첩되었다.

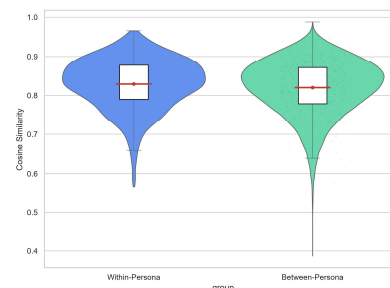


그림 4. 교차 회차 동일·상이 페르소나 응답 유사도 분포

Fig. 4. Cross-run similarity distributions within and between personasparring virtual patient response characteristics

동일 페르소나 검색 성능은 표 6과 같다. Top-1, Top-3 및 Top-5 정확도는 각각 3.41%, 8.18%, 12.73%였으며, 무작위 기대수준은 각각 2.27%, 6.82%, 11.36%였다. 평균 역순위(MRR)는 0.1141, 정답 페르소나의 평균 순위는 20.97이었다. 이러한 결과에서 검색 성능은 무작위 수준을 소폭 상회하였으나, 개별 페르소나 간 구별성은 제한적으로 나타났다.

표 6. 페르소나 구별성 및 검색 성능

Table 6. Persona distinctiveness and retrieval performance

Analysis	Measure	Observed value	Reference value	Difference
Similarity	Within-persona mean	0.8289	–	–
	Between-persona mean	0.8189	–	–
	Within-between gap	0.0100	0	+0.0100
Retrieval	Top-1 accuracy	3.41%	2.27%	+1.14%p
	Top-3 accuracy	8.18%	6.82%	+1.36%p
	Top-5 accuracy	12.73%	11.36%	+1.37%p
	Mean reciprocal rank	0.1141	–	–
	Mean correct rank	20.97	22.50	–1.53

7개 TCI 차원이 모두 정의된 32개 복합 페르소나의 2·3회차 응답 640개를 대상으로 분석한 결과, 위험회피(HA)의 balanced accuracy는 0.5604로 permutation baseline 0.4921보다 높았으며, Holm 보정 후에도 통계적으로 유의하였다($p=.0490$). 반면 다른 TCI 차원에서는 보정 후 유의한 차이가 확인되지 않았다. 따라서 TCI 차원 수준에서는 일부 설정과 관련된 텍스트 신호가 나타났으나, 이러한 차이가 차원 전반에서 일관되게 확인되지는 않았다.

V. 논의 및 한계

동일한 생성조건이 확인된 2회차와 3회차에서 반복 생성 응답의 핵심 의미가 대체로 유지된 것은 동일한 환자정보와 페르소나 조건, 우울증이라는 공통 임상 조건이 응답의 전반적인 방향을 제한했기 때문으로 해석할 수 있다. 다만 생성형 언어모델의 특성상 표현의 차이 자체보다 반복 생성 과정에서 유지되어야 할 임상 사실이 달라지는 경우가 더 중요하다. 실제로 불일치는 현재 약물 복용 여부를 묻는 Q3과 충동적 행동을 묻는 Q10에 집중되어, 반복 생성에서도 명시적인 임상 사실이나 행동 여부가 항상 동일하게 유지되는 것은 아님을 확인하였다. 이러한 결과는 반복 생성 동안 유지되어야 할 정보가 페르소나 설계 단계에서 충분히 고정되지 않았기 때문으로 볼 수 있다. 따라서 향후 시스템에서는 병력, 치료 경험 및 복용 약물과 같은 고정 임상정보, 기질 및 성격에 따른 반응 특성, 말투와 감정 표현 등의 응답 방식, 고위험 내용에 적용되는 임상 안전 제약을 구분하여 관리하고, 생성 응답이 고정 임상정보와 충돌할 경우 이를 탐지할 수 있는 상태관리 구조를 마련할 필요가 있다. 한편 사후 데이터 무결성 점검에서 1회차의 API 요청에 환자 기본정보가 정상적으로 반영되지 않은 것이 확인되어 주 분석은 2회차와 3회차의 응답으로 제한하였으며, 이에 따라 본 연구의 반복 일관성 평가는 두 번의 유효한 반복 생성에 기반한다는 한계가 있다.

페르소나 간 응답 차이가 크지 않았던 것은 44개의 TCI 조합이 세분화되면서 일부 페르소나의 특성이 유사해졌고, 각 TCI 차원을 구체적인 행동이나 언어 표현과 직접 연결하는 규칙이 별도로 설정되지 않았기 때문으로 해석할 수 있다. TCI 차원 수준의 탐색적 분류에서도 위험회피(HA)에서만 약한 식별 신호가 확인되었고, 다른 차원에서는 다중검정 보정 후 유의한 차이가 나타나지 않았다. 또한 모든 페르소나가 동일한 우울 증례와 공통 대화 규칙을 공유했기 때문에, 일부 질문에서는 페르소나 조건보다 공통 임상 특성과 대화 규칙의 영향이 더 크게 나타났을 가능성이 있다. 따라서 공통 대화 규칙과 TCI 조건의 영향을 분리하여 평가하지 못한 점은 본 연구에서 관찰된 응답 차이를 TCI 조건의 독립적인 효과로 귀속하는 데 있어 내적 타당도의 제한으로 작용한다. 특히 공통 대화 규칙은 응답의 관점과 표현 범위를 일정하게 제한하므로 감정 표현이나 응답 구조와 관련된 일부 분석 지표에도 영향을 미쳤을 수 있다. 따라서 본 연구의 지표는 TCI 특성을 독립적으로 검증하기 위한 척도라기보다 생성 응답의 표현 양상을 비교·검토하기 위한 기술적 지표로 해석해야 하며, 키워드 기반 산출 방식 역시 부정 표현, 문맥 및 다의어를 충분히 구분하지 못한다는 한계가 있다. 이에 본 연구의 결과를 TCI의 심리학적 구성개념이 생성 응답에 타당하게 구현되었음을 입증하는 근거로 해석하기는 어렵다. 향후에는 TCI 차원의 각 수준을 관찰 가능한 행동 및 언어 표현과 연결하는 behavioral anchor를 사전에 정의하고, 질문별 반응 방식과

행동 규칙을 구체화하여 해당 설계조건이 생성 응답에 일관되게 나타나는지를 별도로 검증할 필요가 있다.

실제 임상교육에 적용하기 위해서는 생성 응답의 임상적 적절성과 안전성에 대한 추가 검증이 필요하다. 특히 자살사고와 같은 고위험 문항은 응답 내용의 임상적 타당성과 교육 과정에서 부적절하거나 위험한 정보를 제공할 가능성을 임상전문가가 평가할 필요가 있다. 본 연구는 단일 우울 증례와 단일 생성 모델을 대상으로 하였으며, 임상전문가에 의한 내용타당도·안전성 평가와 교수자 및 학습자를 대상으로 한 교육 효과 및 사용자 평가를 수행하지 않았다. 따라서 향후에는 임상전문가를 통한 내용 및 안전성 검증과 함께 교수자와 학습자를 대상으로 시스템의 사용성, 검토 효율, 페르소나 차이의 인지 가능성 및 실제 교육환경에서의 활용 가능성을 평가할 필요가 있다. 또한 실제 임상자료를 활용할 경우에는 개인 식별정보의 비식별화, 입력정보의 최소화, 데이터 접근 및 관리 등 개인정보 보호 절차가 함께 고려되어야 한다.

Gemini 기반 의미 판정 역시 단일 평가모델과 임상전문가가 아닌 단일 연구자의 검토에 기반하였다는 한계가 있다. 평가 과정에서 SBERT 유사도와 연구자 판정을 Gemini에 제공하지 않았으나, 응답의 제시 순서를 교환하지 않았고 동일 입력에 대한 반복 판정 안정성도 별도로 검증하지 않았다. 또한 응답 생성에 사용한 gpt-4.1-mini와 의미 판정에 사용한 gemini-3.5-flash-lite의 세부 모델 snapshot을 기록하지 않았고 일부 생성 및 평가 파라미터를 별도로 지정하지 않아, 동일한 모델 alias를 사용하더라도 결과를 완전히 재현하는 데에는 제한이 있다. 따라서 Gemini 판정은 객관적 정답이나 임상적 타당성 검증 결과라기보다, SBERT 유사도만으로 확인하기 어려운 핵심 의미와 사실 충돌을 보완적으로 검토한 결과로 해석할 필요가 있다. 향후에는 고정된 모델 버전과 생성 조건을 기록하고, 복수의 임상전문가에 의한 독립적 판정과 평가자 간 신뢰도 검증을 통해 평가 절차의 재현성과 임상적 타당성을 보완할 필요가 있다.

VI. 결 론

연구는 임상 교육용 가상환자의 페르소나를 TCI의 기질 및 성격 차원에 따라 구성하고 생성 응답을 비교·검토할 수 있도록, TCI 기반 페르소나 설계와 응답 생성·분석·시각화 과정을 통합한 시스템을 제안하였다. 동일한 생성조건이 확인된 2회차와 3회차에서 기질 및 성격 조합을 달리한 페르소나의 반복 생성 응답을 분석한 결과, 핵심 의미는 대체로 유지되었으나 세부 정보의 완전한 재현은 제한적이었으며, 약물 복용 여부와 같이 사실성이 중요한 질문에서는 응답 충돌이 나타났다. 또한 동일 페르소나 응답이 서로 다른 페르소나 응답보다 다소 유사했지만, 개별 페르소나를 안정적으로 구별할 정도의 차이는 확인되지 않았다.

본 연구는 TCI의 심리학적 구성개념이 LLM의 언어·행동 특성으로 타당하게 구현되었음을 입증한 것이 아니라, TCI 차원을 페르소나 설계조건으로 활용한 시스템을 개발하고 이에 따른 생성 응답의 반복 일관성과 구별성을 예비적·탐색적으로 평가한 연구이다. 분석 결과는 TCI의 수준값을 프롬프트 조건으로 제시하는 것만으로 가상환자의 충분한 페르소나 구별성과 고정 임상정보의 일관성이 자동으로 확보되지는 않음을 보여준다. 따라서 향후에는 TCI 차원별 반응 특성을 관찰 가능한 행동 및 언어 표현과 연결하는 규칙을 구체화하고, 성격적 반응 경향과 병력 및 약물 복용 여부 등의 고정 임상정보를 분리하여 관리하며, 응답 방식과 임상 안전 제약을 함께 고려하는 상태관리 구조를 마련할 필요가 있다. 제안 시스템은 페르소나의 심리학적 타당성을 자동으로 보장하는 도구라기보다, 교수자가 이론적 설계 기준에 따라 가상환자를 구성하고 생성 결과를 체계적으로 비교·검토할 수 있도록 지원한다는 데 의의가 있다. 향후에는 임상전문가를 통한 생성 내용의 타당도와 안전성 평가와 함께, 임상 교육자와 학습자를 대상으로 시스템의 사용성, 검토 효율 및 페르소나 차이의 인지 가능성을 검증하여 실제 교육 환경에서의 활용 가능성을 확인할 필요가 있다.

감사의 글

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 지역진흥화혁신인재양성사업(IITP-2026-RS-2020-II201791, 50%), 정부(과학기술정보통신부)의 재원으로 한국연구재단(RS-2025-00519817, 50%)과 부산광역시 및 (재)부산테크노파크의 BB21plus 사업으로 지원된 연구임.

참고문헌

- [1] R. M. Epstein and E. M. Hundert, "Defining and Assessing Professional Competence," *JAMA*, Vol. 287, No. 2, pp. 226-235, January 2002. <https://doi.org/10.1001/jama.287.2.226>
- [2] W. C. McGaghie, S. B. Issenberg, E. R. Cohen, J. H. Barsuk, and D. B. Wayne, "Does Simulation-Based Medical Education with Deliberate Practice Yield Better Results than Traditional Clinical Education? A Meta-Analytic Comparative Review of the Evidence," *Academic Medicine*, Vol. 86, No. 6, pp. 706-711, June 2011. <https://doi.org/10.1097/ACM.0b013e318217e119>
- [3] S. B. Issenberg, W. C. McGaghie, E. R. Petrusa, D. Lee Gordon, and R. J. Scalese, "Features and Uses of High-Fidelity Medical Simulations That Lead to Effective

Learning: A BEME Systematic Review," *Medical Teacher*, Vol. 27, No. 1, pp. 10-28, January 2005. <https://doi.org/10.1080/01421590500046924>

- [4] A. Gélis, S. Cervello, R. Rey, G. Llorca, P. Lambert, N. Franck, ... and B. Rolland, "Peer Role-Play for Training Communication Skills in Medical Students: A Systematic Review," *Simulation in Healthcare*, Vol. 15, No. 2, pp. 106-111, April 2020. <https://doi.org/10.1097/SIH.00000000000000412>
- [5] H. S. Barrows, "An Overview of the Uses of Standardized Patients for Teaching and Evaluating Clinical Skills," *Academic Medicine*, Vol. 68, No. 6, pp. 443-451, June 1993. <https://doi.org/10.1097/00001888-199306000-00002>
- [6] H. M. Bosse, M. Nickel, S. Huwendiek, J. H. Schultz, and C. Nikendei, "Cost-Effectiveness of Peer Role Play and Standardized Patients in Undergraduate Communication Training," *BMC Medical Education*, Vol. 15, 183, October 2015. <https://doi.org/10.1186/s12909-015-0468-1>
- [7] J. Zeng, W. Qi, S. Shen, X. Liu, S. Li, B. Wang, ... and S. Cao, "Embracing the Future of Medical Education with Large Language Model-Based Virtual Patients: Scoping Review," *Journal of Medical Internet Research*, Vol. 27, e79091, November 2025. <https://doi.org/10.2196/79091>
- [8] F. Holderried, C. Stegemann-Philipps, A. Herrmann-Werner, T. Festl-Wietek, M. Holderried, C. Eickhoff, and M. Mahling, "A Language Model-Powered Simulated Patient with Automated Feedback for History Taking: Prospective Study," *JMIR Medical Education*, Vol. 10, e59213, August 2024. <https://doi.org/10.2196/59213>
- [9] R. Scherr, F. F. Halasch, A. Spina, S. Andalib, and R. Rivera, "ChatGPT Interactive Medical Simulations for Early Clinical Education: Case Study," *JMIR Medical Education*, Vol. 9, e49877, November 2023. <https://doi.org/10.2196/49877>
- [10] D. Weisman, A. Sugarman, Y. M. Huang, L. Gelberg, P. A. Ganz, and W. S. Comulada, "Development of A GPT-4-Powered Virtual Simulated Patient and Communication Training Platform for Medical Students to Practice Discussing Abnormal Mammogram Results with Patients: Multiphase Study," *JMIR Formative Research*, Vol. 9, e65670, April 2025. <https://doi.org/10.2196/65670>
- [11] S. Maaz, J. C. Palaganas, G. Palaganas, and M. Bajwa, "A Guide to Prompt Design: Foundations and Applications for Healthcare Simulationists," *Frontiers in Medicine*, Vol. 11, 1504532, January 2025. <https://doi.org/10.3389/fmed.2024.1504532>
- [12] C. R. Cloninger, D. M. Svrakic, and T. R. Przybeck, "A Psychobiological Model of Temperament and Character,"

- Archives of General Psychiatry*, Vol. 50, No. 12, pp. 975-990, December 1993. <https://doi.org/10.1001/archpsyc.1993.01820240059008>
- [13] A. Ziv, P. R. Wolpe, S. D. Small, and S. Glick, "Simulation-Based Medical Education: An Ethical Imperative," *Academic Medicine*, Vol. 78, No. 8, pp. 783-788, August 2003. <https://doi.org/10.1097/00001888-200308000-00006>
- [14] A. H. Al-Elq, "Simulation-Based Medical Teaching and Learning," *Journal of Family and Community Medicine*, Vol. 17, No. 1, pp. 35-40, January 2010. <https://doi.org/10.4103/1319-1683.68787>
- [15] C. Lane and S. Rollnick, "The Use of Simulated Patients and Role-Play in Communication Skills Training: A Review of the Literature to August 2005," *Patient Education and Counseling*, Vol. 67, No. 1-2, pp. 13-20, June 2007. <https://doi.org/10.1016/j.pec.2007.02.011>
- [16] H. S. Barrows and S. Abrahamson, "The Programmed Patient: A Technique for Appraising Student Performance In Clinical Neurology," *Journal of Medical Education*, Vol. 39, pp. 802-805, August 1964.
- [17] N. V. Vu and H. S. Barrows, "Use of Standardized Patients in Clinical Assessments: Recent Developments and Measurement Findings," *Educational Researcher*, Vol. 23, No. 3, pp. 23-30, April 1994. <https://doi.org/10.3102/0013189X023003023>
- [18] C. Gillette, R. B. Stanton, N. Rockich-Winston, M. Rudolph, and H. G. Anderson Jr., "Cost-Effectiveness of Using Standardized Patients to Assess Student-Pharmacist Communication Skills," *American Journal of Pharmaceutical Education*, Vol. 81, No. 10, 6120, December 2017. <https://doi.org/10.5688/ajpe6120>
- [19] A. Hamilton, A. Molzahn, and K. McLemore, "The Evolution from Standardized to Virtual Patients in Medical Education," *Cureus*, Vol. 16, No. 10, e71224, October 2024. <https://doi.org/10.7759/cureus.71224>
- [20] D. A. Cook, J. Overgaard, V. S. Pankratz, G. Del Fiore, and C. A. Aakre, "Virtual Patients Using Large Language Models: Scalable, Contextualized Simulation of Clinician-Patient Dialogue with Feedback," *Journal of Medical Internet Research*, Vol. 27, e68486, April 2025. <https://doi.org/10.2196/68486>
- [21] J. Zhuo, S. Zhang, X. Fang, H. Duan, D. Lin, and K. Chen, "ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami: FL, pp. 1950-1976, November 2024. <https://doi.org/10.18653/v1/2024.findings-emnlp.108>
- [22] A. Chatterjee, H. S. V. N. S. K. Renduchintala, S. Bhatia, and T. Chakraborty, "POSIX: A Prompt Sensitivity Index for Large Language Models," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami: FL, pp. 14550-14565, November 2024. <https://doi.org/10.18653/v1/2024.findings-emnlp.852>
- [23] M. A. Oldham, "Personality-Informed Care: Speaking the Language of Personality," *Psychosomatics*, Vol. 61, No. 3, pp. 220-230, May-June 2020. <https://doi.org/10.1016/j.psym.2020.01.005>
- [24] M. Lutz, I. Sen, G. Ahnert, E. Rogers, and M. Strohmaier, "The Prompt Makes the Person(a): A Systematic Evaluation of Sociodemographic Persona Prompting for Large Language Models," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou, China, pp. 23212-23237, November 2025. <https://doi.org/10.18653/v1/2025.findings-emnlp.1261>
- [25] J. M. Hwang and K. Chae, "The Effects of Temperament and Character on Depression in Female College Students," *Korean Journal of Health Psychology*, Vol. 12, No. 4, pp. 969-982, 2007. <https://doi.org/10.17315/kjhp.2007.12.4.018>
- [26] Y. H. Kim, B. J. Ham, M. S. Lee, and H.-Y. Lee, "Temperament and Character Inventory and Remission State of Depression," *Journal of the Korean Neuropsychiatric Association*, Vol. 42, No. 3, pp. 335-339, May 2003.
- [27] Y. K. Lim, Y. M. Choi, and J. Y. Choi, "The Relationship among Temperament, Character, Coping Style, and Depression in Depressive Patients," *Korean Journal of Clinical Psychology*, Vol. 32, No. 2, pp. 331-350, May 2013. <https://doi.org/10.15842/kjcp.2013.32.2.004>
- [28] J. Y. Choi, "Clinical Features According to Temperament Clusters in Depressive Disorder," *Korean Journal of Clinical Psychology*, Vol. 35, No. 4, pp. 711-722, 2016. <https://doi.org/10.15842/kjcp.2016.35.4.001>
- [29] Korea Patient Safety Reporting & Learning System. Suicide Risk Screening Guidance [Internet]. Available: <https://www.kops.or.kr/portal/ifm/infoProvdstdrDetail.do?infoProvdstdrNo=22&searchInfoProvdstdrSe=RLM311001>.
- [30] The Korean Society of Oriental Neuropsychiatry. Clinical Practice Guideline of Korean Medicine for Depression [Internet]. Available: https://nikom.or.kr/nckm/module/practiceGuide/view.do?guide_idx=337&menu_idx=14.



최영미(Young-Mi Choi)

2025년 : 동의대학교 학사

2025년~현 재: 동의대학교 인공지능학과 석사과정
※관심분야 : AI, HCI, 인터페이스, 데이터 시각화 등



허수진(Sujin Hur)

2021년 : 동의대학교 학사
2024년 : 동의대학교 석사

2025년~현 재: 동의대학교 인공지능학과 박사과정
※관심분야 : Edutech, Data Analytics, AI, Chatbot 등



김성희(Sung-Hee Kim)

2006년 : 이화여자대학교 학사
2008년 : 이화여자대학교 석사
2024년 : Purdue University Ph.D

2015년~2017년: 삼성 전자 소프트웨어센터 책임연구원
2019년~2023년: 동의대학교 빅데이터인공지능센터 소장
2017년~현 재: 동의대학교 인공지능학과 부교수
2024년~현 재: 동의대학교 인공지능 대학원 주임교수
※관심분야 : HCI, User-centered AI, Data Visualization 등