

신뢰 지향 아바타 상담봇의 설계와 신호 동의 응답

스무 개 신뢰 신호를 심고, 하나씩 동의를 얻었는지 물었다

Computers and Education: Artificial Intelligence (Elsevier) 투고 준비 원고 · 전체 검토용 한글판

※ 실제 투고본은 영문입니다. 본 문서는 내용 검토를 위한 한글 대역본입니다.

연구 요약

- 스무 개의 신뢰 신호를 다섯 계층으로 나누어 하나의 상담봇에 구현하였다
- 전공이 학생에 대해 아는 것을 답변으로 옮기는 배경 인식 계층을 더하였다
- 각 신호에 해당되는 응답자의 83.3%에서 100%가 동의하였다
- 보수적 판정 기준(0.80)을 통과한 것은 스무 개 가운데 여섯 개이다
- 전원이 신뢰를 보고하였으나 개별 신호는 편차를 보여, 동의 응답과 신뢰는 별개이다

요약

대화형 에이전트의 신뢰 관련 기제는 대체로 하나씩 따로 검증되어 왔고, 설계자가 심은 단서가 정작 그 대상에게 동의를 얻는지는 거의 묻지 않았다. 본 연구는 국제 학부생을 위한 멀티모달 아바타 상담봇에 신뢰 지향 설계를 적용하고, 그 설계가 사용자에게서 얼마나 동의를 얻었는지를 신호 단위로 측정하였다. 챗봇 정체성·답변 품질·대화 자연스러움·운영 정책 투명성, 그리고 전공이 학생에 대해 아는 것을 답변으로 옮기는 배경 인식 계층에 걸쳐 스무 개의 신뢰 신호를 규정하고, 각 신호에 근거와 구현 위치와 측정 문항을 함께 묶었다. 시스템은 검색 증강 생성과 아바타를 결합하며 대면형·음성·텍스트 세 모드를 하나의 백엔드에서 운영한다. 9일간 배포하여 응답자 39 명으로 평가하였다. 스무 개 신호 각각이 해당되는 응답자의 83.3%에서 100% 사이에서 보고되었고 다섯 계층은 서로 6 퍼센트포인트 안에 놓였다. 동의 응답이 눈에 띄는 신호에 물리면서 나머지를 희생시키지 않았다는 뜻이다. 월슨 구간의 하한이 0.80 을 넘어야 한다는 보수적 규칙에서는 여섯 개가 설계 목표를 충족하였으며, 이 표본 크기에서 나머지 열네 개는 어느 방향으로도 판정되지 않는다. 전반 신뢰는 모든 응답자가 보고한 반면 개별 신호는 편차를 보였고, 이는 총합 신뢰 판단을 부분들이 작동한다는 근거로 사용할 수 없음을 보여준다.

주제어: 신뢰 지향 설계, 대화형 에이전트, 아바타, 검색 증강 생성, 국제학생, 신호 동의 응답

1. 서론

1.1 연구 배경

대화형 에이전트는 고등교육의 여러 기능에 적용되어 왔으며, 66 편을 다룬 스코핑 리뷰는 그 활용 범위와 함께 지금까지 보고된 이점과 과제를 정리하고 있다 (Pereira et al., 2023). 이후 대규모 언어모델은 상담 보조 시스템의 역량을 크게 끌어올렸다. 최근의 한 시스템은 기관 문서에 대한 검색과 이메일 기반의 사람 개입을 결합해 학생이 학칙을 찾아가도록 돕고 (Tamascelli et al., 2025), 북미·유럽·아시아·아프리카 대학 사례들은 학업 상담, 개인화 학습, 행정 지원에 챗봇이 쓰이고 있음을 보고한다 (Kalashi et al., 2025).

학업 상담은 이러한 지원의 자연스러운 적용 대상이다. 상담에 대한 학생의 접근은 고르지 않으며, Soria(2023)는 코로나 19 시기 학부생의 학업 상담 접근에 격차가 있었음을 보고했다. 상담자 역시 부담을 받고 있어, Soria 외(2023)는 학업 상담자의 소진을 기록했다. 상시 이용 가능한 보조 시스템이 이 두 조건을 그 자체로 해소하지는 못하지만, 일상적인 상담 질문에 답하는 시간대와 처리량을 넓힐 수는 있다.

학생이 그러한 시스템을 실제로 이용하는지는 그것을 신뢰하는지에 달려 있다. 신뢰는 사람이 자동화 시스템에 의존할지를 좌우하며, 특히 복잡성이나 예상하지 못한 상황 때문에 시스템을 완전히 이해하기 어려울 때 그러하다 (Lee and See, 2004). 인공지능에 대한 신뢰는 인지적 기반과 정서적 기반 모두에 놓이며, AI가 어떤 형태로 표상되는지가 두 신뢰 모두의 선행 요인이다 (Glikson and Woolley, 2020). 교육 맥락에서는 Ayanwale 과 Molefi(2024)가 지각된 신뢰에서 학부생의 챗봇 이용 의도로 이어지는 직접 경로를 보고했고, Polyportis 와 Pahos(2025)는 신뢰를 대규모 언어모델 챗봇에 대한 학생 태도의 선행 요인으로 두었다.

이러한 고려는 국제학생에게 더 첨예해진다. 이들은 언어와 제도의 경계를 넘어 대학 서비스에 접근해야 한다. 언어와 제도의 장벽이 서비스 접근을 가로막는다는 것은 여러 환경에서 기록되어 있으며, 그 어려움은 재학 기간이 늘어난다고 해서 반드시 해소되지 않는다 (Hiratsuka, 2016; Newton et al., 2021; Tajvar et al., 2024). 이 근거는 2.4 절에서 다시 다룬다. 이 집단은 이미 대화형 AI를 비공식적인 1차 지원 수단으로 사용하고 있다 (Nourian et al., 2026).

두 가지 기술적 진전이 신뢰 지향 상담 에이전트를 가능하게 한다. 검색 증강 생성은 모델을 매개변수 기억에만 의존시키지 않고 외부 색인에서 가져온 구절에 조건화하여, 더 구체적이고 다양하며 사실적인 언어를 생성한다 (Lewis et al., 2020). 또한 검색을 대화 루프에 두면 지식 기반 대화에서 환각이 감소한다 (Shuster et al., 2021). 에이전트 표상은 사회적 실재감을 높인다. 에이전트 시각 표상의 의인화 수준을 높이면 지각된 사회적 실재감이 유의하게 증가했으며 (Li et al., 2023), 인간형 임베디먼트와 사람 음성은 사회적 실재감을 높이고 이것이 다시 신뢰 믿음을 증진시켰다 (Qiu and Benbasat, 2009). 그러나 실제로 이러한 기제들은 대개 하나씩 도입되고 평가된다. 이들을 하나의 시스템에 설계로 통합할 수 있는지, 그리고 의도된 각 신뢰 신호가 실제로 사용자에게서 동의 응답을 받는지는 여전히 불분명하다.

1.2 연구 문제와 연구 질문

선행 연구는 신뢰 관련 기제들을 대체로 하나씩 검증해 왔다. 검색 증강은 사실 그라운드팅을 중심으로 (Lewis et al., 2020; Shuster et al., 2021), 에이전트 표상은 사회적 실재감에 미치는 효과를 중심으로 (Qiu and Benbasat, 2009; Li et al., 2023; Liew et al., 2017), 챗봇 신뢰는 그 구조와 예측 요인을 중심으로 다루어졌다 (Glikson and Woolley, 2020; Ng and Zhang, 2025). 각 기제가 개별적으로 기여한다는 사실이, 이들을 결합한 시스템이 전체로서 신뢰받는다는 것을 보장하지는 않는다. 여기서 두 가지 문제가 따라온다. 첫째는 이러한 기제들을 임기응변으로 쌓지 않고 일관된 설계로 조직·구현할 수 있는가이다. 둘째는 구현된 각 신호가 실제로 사용자에게서 동의 응답을 받는가이다.

두 번째 문제에는 신뢰성 연구의 근거가 있다. 현저성-해석 이론에 따르면 사용자는 어떤 요소를 먼저 알아차린 뒤에야 그것을 평가하므로, 눈에 띄지 않은 요소는 판단에 들어오지 못한다 (Fogg, 2003). 신뢰 실험은 안심 단서로 의도적으로 배치되었음에도 사용자에게 자주 간과되거나 잘못 읽혔고 (Kirlappos et al., 2012), 이 이론을 실증적으로 검증한 연구는 현저성과 해석 중 어느 하나가 아니라 둘의 결합이 웹사이트 신뢰성 평가를 구분함을 확인했다 (George et al., 2016). 이 결과들은 대화형 에이전트가 아니라 웹사이트 신뢰성 연구에서 나온 것이므로, 본 연구는 이를 에이전트에 대한 확립된 결과가 아니라 설계 원리로 차용한다. 그럼에도 이는 총합 신뢰 점수에서 추론하는 대신 신호별로 동의 응답을 측정할 근거가 된다. 본 논문에서 동의 응답은 응답 가능한 응답자 가운데 그 신호가 존재했고 의도한 성질을 지녔다는 데 동의한 비율을 뜻한다. 동의 응답은 인지와 같지 않으며, 그 이유는 4.1 절에 밝힌다.

국제학생 상담은 세 번째 문제를 더한다. 챗봇 정체성, 답변 품질, 대화 자연스러움, 운영 정책 투명성을 넘어, 이들 사용자는 시스템이 자신의 배경과 이전 방문, 처한 상황을 알아보는지에 민감할 수 있다. 이 지점의 근거는 얇다. 개인화는 소규모 연구에서 에이전트에 대한 신뢰와 연관된 바 있는데, 여기서는 사용 중 측정된 신뢰·작업부하·수행에 따라 설명이 한 세션 안에서 조정되었다 (Verhagen et al., 2023). 또한 개인화는 환자의 증상 추이와 초기 신뢰에 따라 세션 사이에 조정되는 동기부여 메시지 설계에도 적용되었다 (Tielman et al., 2019). 두 경우 모두 시스템은 이용 중에 스스로 측정한 상태에 적응한다. 사용자의 가입 시 배경 정보를 세션을 넘어 이어받아 알아보는 것을 독립된 신뢰 계층으로 규정하거나, 그러한 인식이 사용자에게서 동의 응답을 받는지를 측정한 연구는 저자들이 아는 한 없다.

동일한 설계 접근을 국내 환경에 배포한 선행 연구는 미묘한 단서가 명시적 단서보다 덜 인지된다고 보고했다 (Authors, 2026). 그 패턴이 다른 코호트에서도 재현되는지는 검증되지 않았다. 이 분야에서 재현 연구는 드물며 (Hornbæk et al., 2014; Wilson et al., 2011, 2012, 2013, 2014), 그로부터 따라 나오는 함의는 2.5 절에서 다룬다.

이에 본 연구는 세 가지 연구 질문을 설정한다. RQ1은 시스템 구축에 관한 것이고, RQ2는 그 구축이 사용자에게 얼마나 동의 응답을 받는지에 관한 것이며, RQ3는 그 동의 응답이 전반 신뢰와 부합하는지에 관한 것으로, 세 질문을 함께 검증한다.

- **RQ1.** 배경 인식 계층을 더한 신뢰 지향 아바타 상담봇을 국제 대학생용으로 설계하고 배포할 수 있는가? - **RQ2.** 각 신뢰 신호는 어느 정도로 사용자의 동의 응답을 받으며, 요소 기반 신호와 동작 기반 신호

사이에 체계적 차이가 있는가? - **RQ3.** 신호별 동의 응답은 전반 신뢰와 어떻게 연결되며, 배경 인식 신호는 얼마나 동의 응답을 받는가?

RQ2와 RQ3는 탐색적 횡단 설문 기술 통계 분석으로 다루며, 변수 간 인과 관계 추론은 후속 연구의 과제로 남긴다. 사례 도메인은 경영 계열 국제 학부생을 대상으로 한 전공 소개와 학업 상담이다.

1.3 연구의 기여 및 논문 구성

본 연구는 신뢰 지향 접근을 멀티모달 아바타 상담봇에 구현하고 그 신호들이 사용자에게 얼마나 동의 응답을 받는지를 검증한다. 본 연구는 분석 체계, 시스템 아키텍처, 실증 사례, 그리고 동의 응답과 신뢰의 관계에 관한 발견이라는 네 측면에서 기여한다.

첫째, 챗봇 정체성, 답변 품질, 대화 자연스러움, 운영 정책 투명성, 배경 인식에 걸친 5개 계층 20개 신뢰 신호로 구성된 신뢰 지향 설계 체계를 제안한다. 개인화는 시스템이 이용 중에 측정한 상태에 적응하는 방식으로 연구되어 왔으나, 가입 시 기록된 사용자의 배경을 세션을 넘어 이어받아 다시 드러내는 것을 독립된 신뢰 계층으로 규정하고 측정한 연구는 없었으며, 본 연구가 이를 처음으로 수행한다.

둘째, 검색 증강 생성과 임베디드 아바타를 결합하고, 대면형·음성·텍스트 세 모달리티를 단일 백엔드에서 운영하며, 사용자 데이터를 외부 서비스가 아닌 기관 서버에 보관하는 다계층 아키텍처의 설계와 구현을 보고한다.

셋째, 국제 학부생을 대상으로 한 실증 배포를 보고하고, 구현된 모든 신호에 대한 신호별 동의율을 제시한다. 동일한 설계 접근과 동일한 측정 절차가 앞선 국내 배포에 적용되었으므로 (Authors, 2026), 명시적 단서와 미묘한 단서 사이의 격차가 다른 코호트에서도 재현되는지를 함께 검증하며, 이는 이 분야에 재현 연구가 드물다는 문제에 대응한다.

넷째, 신호별 동의 응답과 전반 신뢰가 함께 움직이지 않는다는 점을 보고한다. 여러 신호가 다른 신호에 비해 뚜렷하게 적은 수의 사용자에게만 동의 응답을 받은 반면, 전반 신뢰는 모든 응답자가 보고하였다. 따라서 설계된 신호에 동의하는 것과 시스템을 신뢰하는 것은 서로 구별되는 결과로 보이며, 이는 신뢰 지향 시스템을 어떻게 평가할 것인가에 함의를 가진다.

본 논문의 구성은 다음과 같다. 2장은 챗봇 신뢰, 검색 증강 생성, 에이전트 표상, 국제학생 대상 AI에 관한 선행 연구를 정리한다. 3장은 시스템 설계와 5계층 20신호 체계를 기술한다. 4장은 설문 도구, 배포, 분석 절차를 기술한다. 5장은 표본, 신호별 동의 응답, 계층 수준 비교, 배경 인식 계층, 자유응답을 보고한다. 6장은 동의 응답 격차, 동의 응답과 신뢰의 관계, 한계, 후속 연구 방향을 논의한다.

2. 선행 연구

2.1 AI와 챗봇에 대한 신뢰

신뢰는 상대가 자신의 이익에 부합하게 행동하리라는 기대를 바탕으로 취약성을 감수하려는 의향으로 이해된다. 자동화 시스템에 적용하면, Lee와 See(2004)는 신뢰를 사람이 자동화에 의존할지를 좌우하는 요인으로 보았으며, 특히 복잡성이나 예상하지 못한 상황 때문에 완전한 이해가 어려울 때 그러하다고 보았다. 이들은 설계 목표가 신뢰의 극대화가 아니라 적절한 의존이라고 주장했다.

이후 인공지능 연구는 그 판단의 근거를 구분해 왔다. Glikson과 Woolley(2020)는 신뢰성·가시성·투명성·즉각적 반응 같은 단서에 기반한 인지적 신뢰를 정서적 신뢰와 구분하고, AI가 어떤 형태로 표상되는지를 두 신뢰 모두의 선행 요인으로 지목했다. Lukyanenko 외(2022)는 AI 신뢰 연구가 공통의 개념적 기반 없이 분절되어 있다고 지적하고, 이에 대응하는 기초 프레임워크를 제안했다.

챗봇에 특화된 리뷰들은 이 분야를 실증적으로 정리했다. Ng와 Zhang(2025)은 신뢰 예측 요인을 사용자·기계·상호작용·사회적·맥락 요인으로 묶고, 신뢰의 결과를 정서적·관계적·행동적·인지적·심리적 영역에 걸쳐 정리했다. 이들은 또한 이 문헌이 횡단 설계에 치우쳐 있다고 지적하며 신뢰의 동태에 관한 종단 연구를 요청했다. Wang 외(2026)는 챗봇 이용자 질적 인터뷰 연구에서 인간다운 단서와 지각된 신뢰성이 모두 신뢰 서술의 주제로 나타남을 확인했다.

교육 맥락의 근거는 더 구체적이지만 동시에 더 이질적이다. Ayanwale와 Molefi(2024)는 상대적 이점·적합성·시험가능성과 더불어 지각된 신뢰에서 학부생의 챗봇 이용 의도로 이어지는 직접 경로를 보고했다. Polyportis와 Pahos(2025)는 신뢰를 의도가 아니라 학생 태도의 선행 요인으로 두었고, Shahzad 외(2024)는 지각된 신뢰를 인지도와 지각된 유용성 사이의 조절 변수로 모형화했다. Bilquise 외(2024)는 두 대학의 학생 207명을 조사해, 학업 상담 챗봇 채택 의도를 이끈 것은 지각된 유용성이나 신뢰가 아니라 지각된 사용 용이성과 사회적 영향이었음을 확인했다. 따라서 신뢰는 이들 모형에서 서로 다른 위치에 놓이며, 모두 지속적 사용이 아니라 의도를 측정한다.

본 연구에는 두 가지 함의가 따른다. 신뢰가 인지적 기반과 정서적 기반 모두에 놓이므로, 신뢰 지향 설계는 사실 정확성과 인터페이스의 대인적 특질을 함께 다루어야 한다. 또한 이들 프레임워크가 신뢰를 사용자가 지각하는 단서의 결과로 기술하므로, 설계된 신호는 인지된 뒤에야 기여할 수 있으며, 이것이 본 연구가 신호별로 동의 응답을 측정하는 이유이다.

2.2 검색 증강 생성과 사실 그라운드

모델 출력을 검색 가능한 출처에 근거시키는 것은 불안정한 생성에 대한 표준적인 기술적 대응이다. Lewis 외(2020)는 검색 증강 생성을 제안했는데, 이는 시퀀스-투-시퀀스 모델을 매개변수 기억에만 의존시키지 않고 외부 비매개변수 색인에서 가져온 구절에 조건화하는 방식이며, 순수 매개변수 기준선보다 더 구체적이고 다양하며 사실적인 언어를 생성한다고 보고되었다. 이어 Shuster 외(2021)는 검색을 대화 루프에 두면 대화 능력을 유지하면서도 지식 기반 대화에서 환각이 크게 감소함을 보였다.

후속 연구는 이 접근을 더 좁은 영역에서 검증해 왔다. Xiong 외(2024)는 의학 분야 검색 증강 생성을 벤치마크했고, Béchard와 Marquez Ayala(2024)는 기업 과제의 구조화 출력 생성에서 환각이 감소함을 보고했다. 따라서 현재 확보된 근거는 도메인에 한정된다. 해당 시기의 대화 모델, 의학 질의응답, 기업 구조화 출력을 포괄할 뿐, 오늘날의 지시 조정 모델 전반에 대해 환각이 감소한다는 일반적 결론을 세우지는 않는다. 상담 에이전트에 대해 유효한 함의는 더 좁고 안전하다. 즉 기관 코퍼스에 연결된 답변은 매개변수 기억만으로 생성된 답변보다 지어낼 가능성이 낮다는 것이다.

본 시스템에서 검색은 2.1 절에서 확인한 신뢰의 인지적 기반에 복무한다. 검색은 신뢰의 충분조건이 아니라 구현된 여러 신뢰 신호 가운데 하나로 다루어진다.

2.3 에이전트 표상과 사회적 실재감

에이전트가 사용자에게 어떤 형태로 표상되는지는 그 존재가 얼마나 실재하는 것처럼 느껴지는지에 영향을 준다. Qiu와 Benbasat(2009)은 실험실 연구에서 인간형 임베디먼트와 사람 음성이 함께 지각된 사회적 실재감을 유의하게 높였고, 이것이 다시 신뢰 믿음과 의사결정 보조로서의 이용 의도를 증진시켰음을 확인했다. Li 외(2023)는 헬스케어 에이전트 시각 표상의 의인화 수준을 조작해 수준이 높을수록 지각된 사회적 실재감이 유의하게 증가함을 보였다. 이들의 경로 모형에서 사회적 실재감과 무의식적 의인화는 신뢰를 예측한 반면, 프라이버시 우려는 신뢰를 부적으로 예측했다. Liew 외(2017)는 전자상거래 사이트에 비상호작용 말하는 아바타를 추가해, 아바타가 없는 동일 사이트에 비해 지각된 사회적 실재감이 개선됨을 확인했다.

세 가지 단서를 달아야 한다. 첫째, 이들 연구에서 반복되는 발견은 표상이 사회적 실재감을 높인다는 것이며, 신뢰에 대한 효과는 직접적이지 않고 매개된다. 둘째, 그 효과가 언제나 긍정적인 것은 아니다. Liew 외(2017)에서 신뢰도와 이용 의사에 대한 효과는 성별에 의해 조절되었고, Kim 외(2024)는 정서적으로 표현하는 가상 에이전트가 집단 토론 과제에서 지각된 사회적 실재감과 신뢰도를 오히려 낮췄으며, 참여자가 가장 선호한 구성은 과제에 관여하되 정서 표현은 하지 않는 에이전트였음을 확인했다. 셋째, 임베디먼트만이 관계적 지렛대인 것은 아니다. Bickmore와 Picard(2005)는 임베디먼트가 아니라 관계적 특질을 조작해, 사회정서적이고 관계 형성적인 대화를 사용하는 에이전트가 동등한 과업 지향 에이전트보다 4 주간의 사용에 걸쳐 더 존중받고 호감을 얻으며 신뢰받았음을 확인했다.

따라서 본 연구는 임베디먼트를 사회적 실재감을 높일 것으로 기대되는 신호로 채택하되, 신뢰는 당연한 귀결이 아니라 하류의 매개된 결과로 다룬다.

2.4 AI와 국제학생

국제학생은 추가적인 언어적·제도적 여과를 거쳐 대학 서비스에 접근한다. Hiratsuka(2016)는 일본어가 국제학생의 캠퍼스 서비스 접근에 조직적 장벽으로 작용한다고 기술했고, Newton 외(2021)는 호주 한 대학의 학생 21 명을 대상으로 한 질적 연구에서 정보적·문화적 장애와 함께 지원 개선의 여지를 확인했다. Perry 외(2020)는 한 대학에서 국제학생과 국내학생의 학생 서비스 인지·이용·가치 평가를 비교해 격차가 서비스의 제공이 아니라 인지에 있음을 밝혔다. 이는 본 연구가 신호 단위로 재는 구분, 곧 무

엇을 넣었는가와 무엇이 사용자에게서 동의를 얻었는가의 구분과 같다. Tajvar 외(2024)는 이란의 한 의과대학에서 국제학생 165 명을 조사해 전체 어려움 점수가 재학 기간에 따라 유의하게 달라짐을 밝혔으며(일원배치 분산분석, $p < 0.001$), 3 년 초과 재학자가 가장 높았고($M = 75.1$) 1 년 미만 재학자가 가장 낮았다($M = 47.0$, $n = 11$). 다만 이 연구는 1 학년을 마친 학생만을 대상으로 하였으므로 신입생과 재학생을 대비한 것은 아니다. 이 연구들은 서비스와 정보에 대한 접근을 폭넓게 다루며, 그중 어느 것도 학업 상담만을 따로 떼어 측정하지는 않는다.

생성형 AI 는 이 공간에 빠르게 진입했다. Zhu 와 Yang(2026)은 국제학생의 문화적응에서 생성형 AI 를 정보적·정서적·사회적 지원 채널로 검토했고, Nourian 외(2026)는 혼합방법 연구에서 미국 내 국제학생이 분절된 지원 생태계 안에서 이미 대화형 AI 를 1 차 지원 수단으로 사용하고 있으며, 동시에 무엇까지 말할지에 대한 경계를 분명히 하고 있음을 보고했다. 더 넓게는, 언어 교육에서의 생성형 AI 리뷰가 실시간 피드백과 개별 학습자 요구에 대한 적응을 통해 이러한 도구를 형평 격차를 좁히는 수단으로 자리매김한다 (Weng and Fu, 2025).

따라서 이 집단은 높은 필요와, 시스템을 신뢰할 수 있는지에 대한 높은 민감성을 함께 지니며, 이는 설계된 신뢰 신호가 사용자에게 동의 응답을 받는지를 시험하기에 까다로운 환경이 된다.

2.5 신뢰 지향 설계 접근의 필요성

앞의 절들은 기제별로 조직된 문헌을 기술한다. 신뢰 연구는 신뢰가 무엇에 근거하는지를 규정하고, 검색 연구는 사실 그라운딩을 다루며, 에이전트 표상 연구는 사회적 실재감을 다루고, 국제학생 연구는 그러한 시스템이 사용될 접근 조건을 기술한다. 빠져 있는 것은 이 기제들을 하나의 설계 안에 계층으로 의도적으로 심은 시스템과, 심어진 각 신호가 실제로 사용자에게 동의 응답을 받는지에 관한 근거이다.

두 가지 공백이 본 연구를 규정한다. 첫째, 개인화는 에이전트 연구에서 신뢰 및 관여와 연결되어 왔으나, 다루어진 적응은 시스템이 이용 중에 측정한 상태에 따라 이루어진다. 한 세션 안에서 측정된 신뢰·작업부하·수행에 따르거나 (Verhagen et al., 2023), 세션 사이에 증상 추이에 따르는 식이다 (Tielman et al., 2019). 에이전트에 대한 신뢰를 가장 직접적으로 측정한 것도 소규모 연구이다 (Verhagen et al., 2023). 사용자의 가입 시 배경 정보를 세션을 넘어 이어받아 다시 사용자에게 드러내는 것을 독립된 신뢰 계층으로 규정하거나 그것이 동의 응답을 받는지를 측정한 연구는 저자들이 아는 한 없다. 둘째, 이 분야에서 재현 연구는 드물다. 인간-컴퓨터 상호작용 분야 논문 조사에서 약 3%만이 재현 연구였으며 (Hornbæk et al., 2014), 학계는 재현 확대를 반복적으로 요청해 왔다 (Wilson et al., 2011, 2012, 2013, 2014). 따라서 한 번의 배포에서 평가된 설계 접근은 다른 환경에 다시 적용되기 전까지는 그 위상이 제한된다.

본 연구는 이 두 가지에 대응한다. 신호를 계층별로 규정한 신뢰 지향 시스템을 구현하고, 선례가 확립되지 않은 배경 인식 계층을 추가하며, 각 신호가 동의 응답을 받는지를 측정하고, 이를 그 접근이 처음 평가된 집단과 다른 집단을 대상으로 한 두 번째 배포에서 수행한다.

3. 시스템 설계

3.1 아키텍처 개요

상담 시스템은 그림 1 과 같이 네 계층으로 구성된다. 첫째 계층은 브라우저 클라이언트로, 상담 인터페이스를 렌더링하고 아바타 세션을 유지하며 음성과 텍스트 입력을 받는다. 둘째 계층은 엣지 프록시로, 클라이언트의 대화 요청을 받아 추론 호스트로 전달하며 대화 상태를 보유하지 않고 아무것도 저장하지 않는다. 셋째 계층은 추론 호스트로, 자체 관리 서버에서 검색 파이프라인과 언어모델을 실행한다. 넷째 계층은 기관 서버로, 사용자를 인증하고 대화 로그를 기록하며 설문 응답을 기관 자체 데이터베이스에 저장한다.

이 분리는 1 장에서 밝힌 설계 의도를 따른 것이다. 학생 식별자, 대화 로그, 설문 기록은 기관 계층에만 기록되므로 외부 서비스 사업자에게 사용자 데이터가 축적되지 않는다. 생성은 상용 API 가 아니라 자체 관리 추론 호스트에서 수행되어 모델과 코퍼스를 기관 통제 아래 둔다. 계층 간 통신은 HTTP 로만 이루어지므로 한 계층 내부의 변경이 다른 계층의 변경을 요구하지 않는다.

3.2 검색 파이프라인

답변은 전공의 공개 정보를 정리한 코퍼스에 근거한다. 코퍼스는 전공 개요, 입학, 교육과정, 진로 트랙, 교수진을 다루는 43 개의 질문-답변 청크로 구성되며, 각 청크는 식별자, 섹션 라벨, 질문, 답변을 가진다.

구축 시 각 청크는 다국어 임베딩 모델로 임베딩되어 청크 텍스트와 함께 저장된다. 질의 시 사용자 메시지는 동일 모델로 임베딩되고, 벡터는 L2 정규화되며, 코퍼스에 대해 코사인 유사도가 계산된다. 점수가 높은 상위 세 개 청크를 검색하고, 0.40 미만의 결과는 폐기한다. 임계값을 넘는 청크가 없으면 프롬프트에 어떤 출처 텍스트도 주입되지 않으며, 시스템은 매개변수 기억에서 답을 만들어내는 대신 해당 정보를 가지고 있지 않다고 말하도록 지시받는다. 이 임계값은 주제를 벗어난 질문에 대해 지어낸 답변을 막는 본 시스템의 주된 장치이며, 3.4 절의 검색 그라운드링 신호와 한계 인정 신호를 실제로 구현하는 부분이다.

생성은 추론 호스트에서 로컬로 서빙되는 공개 가중치 지시조정 모델이 수행한다. 검색과 생성이 모두 기관이 관리하는 하드웨어에서 실행되므로, 학생의 질문은 외부 추론 사업자에게 전송되지 않는다.

3.3 멀티모달 인터페이스

상담 화면은 그림 2 와 같이 영상 상호작용 영역과 대화 영역으로 나뉜다. 영상 영역에는 아바타 무대, 세션 시작·종료 제어, 모드 전환, 아바타 발화를 중단시키는 제어가 배치된다. 대화 영역에는 메시지 스트림, 추천 질문 묶음, 그리고 마이크 제어·텍스트 입력창·전송 동작으로 구성된 입력 도크가 배치된다.

세 가지 상호작용 모드가 동일한 백엔드에서 동작한다. 대면형 모드에서는 학생이 말하고 아바타로부터 음성 답변을 받으며, 음성 모드에서는 음성만으로, 텍스트 모드에서는 키보드만으로 주고받는다. 학생은

한 세션 안에서 모드를 전환할 수 있다. 세 모드 모두 동일한 검색·생성 경로를 지나므로 답변의 사실 내용은 선택한 모드에 좌우되지 않는다. 음성 출력은 합성 전에 정규화되어 약어와 전공명이 일관되게 발음된다.

3.4 5 계층 20 신호 프레임워크

시스템에 구현된 신뢰 신호는 다섯 계층으로 조직된다. 앞의 네 계층은 앞선 국내 배포에서 사용한 구성을 따른다 (Authors, 2026). 챗봇 정체성, 답변 품질, 대화 자연스러움, 운영 정책 투명성이다. 다섯째 계층인 배경 인식은 본 연구에서 도입한다.

근거는 개별 신호가 아니라 계층 수준에서 제시한다. 선행 연구가 확립한 것은 기제이다. 임베디먼트가 사회적 실재감을 높인다는 것, 검색 그라운드링이 근거 없는 생성을 줄인다는 것, 프라이버시 우려가 신뢰를 낮춘다는 것이 그것이다. 특정 임계값이나 특정 인사말, 특정 버튼이 그 기제를 구현하는 올바른 방식이라는 것까지 확립한 것은 아니다. 그것들은 본 연구에서 내린 공학적 선택이며, 각각에 인용을 붙이는 것은 인용된 연구가 보인 바를 잘못 진술하는 일이 된다. 따라서 표 1은 계층별 근거를, 표 2는 각 계층이 이 배포에서 취한 구체적 형태로서의 신호를 제시한다.

근거는 다른 맥락에서 옮겨온 것이며, 그 점을 밝혀둔다. 검색 증강의 근거는 개방 도메인 벤치마크의 사전-LLM 시퀀스 모델에서 나온 것이지, 기관 관련 질문에 답하는 지시학습 어시스턴트에서 나온 것이 아니다. 신뢰 보정에 관한 지침은 대화가 아니라 안전 필수 자동화의 감독 제어에서 나온 것이다. 사회적 실재감에 관한 결과는 실험실 쇼핑 과제, 그리고 참여자가 상호작용을 직접 수행한 것이 아니라 영상으로 관찰한 비네트 연구에서 나온 것이다. 각 기제는 새로운 환경으로 옮겨온 설계 원리로 사용되는 것이지, 이 환경에서 성립함이 밝혀진 결과로 사용되는 것이 아니다.

표를 이렇게 나눈 데서 두 가지가 따라 나온다. 신호는 설계 결정으로 제시되며 이는 실제로 그러하다. 그리고 독자는 시스템의 어느 부분이 published 기제에 기대고 어느 부분이 저자의 판단에 기대는지 구분할 수 있다. 또한 다섯째 계층의 비어 있는 근거 칸은 표의 결함이 아니라 공백의 진술이다. L5의 신호들은 세션 사이로 이어지는 정보에 의존하는데, 그것을 신뢰 신호로 다루거나 측정할 선행 연구를 찾지 못하였다. 각 신호를 측정하는 설문 문항은 부록에 제시한다.

각 신호는 하나의 식별자 아래 세 가지가 묶이도록 규정된다. 그 신호가 왜 중요할 것으로 기대되는지, 시스템의 어느 위치에 구현되는지, 그리고 사용자가 그것에 동의 응답을 했는지를 무엇으로 측정하는지이다. 신호를 평면적인 목록이 아니라 계층으로 묶은 것은, 대화형 에이전트에 대한 사용자 판단이 단일한 품질이 아니라 구별 가능한 소수의 수준으로 수렴한다는 통합 리뷰 (Marconi et al., 2026)를 따른 것이다. 여기서 쓴 묶음은 서술의 편의만은 아니다. 그림 2는 각 신호가 인터페이스의 어디에서 표면화되는지 보여준다. 첫째 계층은 특정 교수자를 닮은 아바타, 기관 정체성 표시, 그리고 이 시스템이 AI라는 명시적 고지를 포괄한다. 둘째 계층은 검색 그라운드링, 한계 인정, 말투, 형식 일관성을 포괄한다. 셋째 계층은 응답 속도 조절, 시스템이 자신의 발화를 다시 받아 적는 것에 대한 방지, 사용자의 발화 중단, 아바타 립싱크와 제스처, 모드 전환을 포괄한다. 넷째 계층은 개인정보 동의 화면, 계정 없이 둘러보기, 인사 형식, 방문 횟수 집계, 음성 정규화를 포괄한다.

다섯째 계층은 이 환경에서 도출된다. 가입 시 시스템은 학생의 국가와 관심 트랙을 수집하며, 이는 답변을 구성할 때 활용되어 배경 인지 신호와 배경 맞춤 신호를 뒷받침한다. 방문일수는 계정별로 집계되며, 서로 다른 방문일이 두 번째가 되는 시점부터 재방문 인사가 표시된다. 첫 방문자에게는 아무것도 표시되지 않는다. 이 계층은 기존 개인화 연구의 확장이 아니라 새로운 것으로 다룬다. 선행 연구의 적응은 시스템이 이용 중에 측정한 상태, 곧 진술된 선호나 관찰된 행동에 기반하는 반면, 여기의 신호들은 사용자가 가입 시 한 번 제공하고 다시 말하지 않은 배경 정보에 의존하기 때문이다.

4. 연구 방법

4.1 설문 도구

이용자 인식은 세션 종료 시점에 브라우저에서 모달 형태로 제시되는 설문지로 수집하였다. 설문지는 네 부분으로 구성된다. 1 부는 응답자의 국가, 관심 트랙, 연령대, 성별, 그리고 선택 응답인 MBTI 유형을 기록한다. 성격 유형 항목은 본 논문의 어떤 분석에도 사용하지 않았다. 2 부는 스무 개의 신뢰 신호 문항을 프레임워크의 계층에 대응하는 다섯 개 제목 아래 묶어 제시한다. 3 부는 전반 신뢰를 묻는 단일 문항을 제시한다. 4 부는 두 개의 자유응답을 제시하는데, 하나는 가장 신뢰가 갔던 지점을, 다른 하나는 가장 신뢰가 가지 않았거나 어색했던 지점을 묻는다.

스물한 개의 폐쇄형 문항은 각각 두 개의 버튼 중 하나로 응답한다. 중간값도, "잘 모르겠다"도, 명시적인 무응답 선택지도 없으며, 해당되는 문항에 모두 답하기 전에는 제출할 수 없다. 이분형을 택한 것은 응답 부담을 낮추어 학생이 세션 직후 설문을 이탈하지 않고 완료하도록 하기 위해서였다. 그 대가로 이 도구는 정도를 기록하지 못한다. 희미하게 인지된 신호와 분명하게 인지된 신호가 동일하게 기록된다. 이 한계는 6 장에서 다시 다룬다.

문항은 방금 마친 세션에 관한 1 인칭 진술문으로 서술되어 있고, 응답자는 그에 대한 동의 여부를 표시한다. 따라서 한 문항은 신호를 인지했다는 사실만을 따로 기록하는 것이 아니라, 그 신호가 존재했고 의도한 성질을 지녔다는 데 동의하는지를 기록한다. 본 논문에서 동의 응답은 응답 가능한 응답자 가운데 동의한 비율을 가리키며, 그런 의미로 읽어야 한다. 이러한 동의는 인지와 평가를 함께 묶는다. 아바타의 발화를 인지했으나 입 모양이 잘 맞지 않았다고 판단한 응답자는, 그 신호에 아예 주의를 두지 않은 응답자와 같은 방식으로 기록된다. 그러므로 이 측정치는 인지에 대한 보수적 추정치가 아니며 그렇게 읽어서도 안 된다. 동의는 묵종으로, 또는 신호가 있었으리라는 추측으로 높아질 수 있고, 신호를 인지했으나 부정적으로 평가한 응답자 때문에 낮아질 수 있다. 두 방향은 상쇄되지 않으며 이 도구는 둘을 분리하지 못한다. 나아가 여러 문항은 인지가 아니라 평가·신호·효과를 기록하도록 서술되어 있다. F4 는 검색된 자료가 제시되었는지가 아니라 답변이 정확하게 느껴졌는지를 묻고, F12 는 모드를 전환할 수 있다는 점을 좋아했는지를 묻고, F14 는 계정 없이 열람하는 것이 마음을 더 편하게 했는지를 묻는다. 이들 문항의 결과는 해당 기능에 대한 동의 응답으로 읽으며, 그 기제가 인지되었다는 근거로 읽지 않는다. 스무 문항 가운데 다섯만이 신호의 존재나 인지를 직접 묻고, 나머지 열다섯은 평가·신호·효과를 물으며 그중 하나는 둘이 섞여 있다. 어느 문항이 어느 쪽인지는 부록 A에 기록한다.

다섯 개 문항은 조건부이다. 두 개는 음성 출력에 관한 것으로 아바타 모드나 음성 모드 이용이 기록된 응답자에게만 제시되고, 한 개는 아바타의 가시적 동작에 관한 것으로 아바타 모드가 기록된 경우에만 제시되며, 두 개는 방문 간 인식에 관한 것으로 서로 다른 방문일이 두 번째가 되는 시점부터 제시된다. 해당되지 않는 문항은 표시되지 않고, 제출 요건에도 포함되지 않으며, 부정 응답이 아니라 결측으로 저장된다. 모드 기록은 브라우저에 보관되므로 그 기록이 비어 있는 응답자는 음성 및 아바타 문항에 응답 가능한 것으로 처리되었다. 따라서 이 세 문항의 응답 가능 집단에는 해당 모드를 실제로 이용하지 않은

학생이 포함되었을 수 있다. 결측 처리 방식은 4.3 절에서 기술하며, 이 대체 처리의 결과는 6 장에서 다룬다.

화면에 표시되는 문항 번호는 해당 응답자가 실제로 보는 문항에 대해 순서대로 부여되므로, 같은 신호가 응답자마다 다른 번호로 나타날 수 있다. 이에 따라 본 논문은 신호를 응답자가 본 번호가 아니라 프레임워크 식별자로 지칭한다. 모든 문항의 문구와 선택지 목록을 포함한 설문지 전문은 부록 A의 표 A.1에 수록한다.

4.2 배포와 참여자

상담 시스템은 국내 한 사립대학의 학부 경영 계열 전공을 위해 배포되었고 전공의 공개 웹 주소를 통해 접근되었다. 자료는 2026년 6월 5일부터 13일까지 수집하였다. 설문 진입 지점은 6월 1일부터 사이트 푸터에 있었으며, 응답은 주로 학생들에게 시스템을 소개하고 이용을 권유한 수업 세션에서 수집되었다. 참여는 자발적이었고 보상은 제공되지 않았다. 응답자는 해당 전공의 국제 학부생이다.

설문지 41 부가 제출되었다. 도구는 각 응답자가 양식에 머문 시간을 기록하였으며, 60 초 미만에 완료된 제출 2 건을 부주의 응답으로 제외하였다. 두 건은 모두 6월 9일에 제출되어 각각 45 초와 46 초였고, 유지된 응답 중 가장 빠른 것은 67 초였으므로 이 배제는 경계선상의 판단에 의존하지 않았다. 분석 표본은 39 부이며 완료 시간의 중위값은 2 분 33 초이다. 그 밖의 배제는 적용하지 않았다.

본 연구는 학과 교육 도구에 대한 개선 의견을 수집하였다. 설문지는 이름·연락처를 비롯한 직접 식별자를 묻지 않았으며, 폐쇄형 문항과 함께 국적, 지망 트랙, 연령대, 학년, 성별, 그리고 선택 항목인 성격 유형을 기록하였다. 기관생명윤리위원회 심의를 거치지 않은 근거는 후미의 윤리 진술에 밝힌다.

배포의 두 가지 특성이 이 표본으로 뒷받침할 수 있는 범위를 제약한다. 이를 한계 절에만 두지 않고 여기에 밝히는 것은, 이들이 5 장의 결과를 어떻게 읽어야 하는지에 직접 관련되기 때문이다.

첫째, 설문을 열기 위해 필요한 최소 대화량이 없었다. 턴 수 조건도, 경과 시간 조건도, 상담 시스템을 실제로 이용했는지에 대한 확인도 없었으므로, 방문자는 세션을 갖지 않은 채 푸터에서 설문에 도달할 수 있었다. 39 건의 응답은 36 개의 서로 다른 브라우저 세션에서 도착했으며, 도구에는 동일 브라우저의 두 번째 제출을 막는 장치가 없었다.

둘째, 시스템의 제시 방식이 수집 기간 중간인 6월 10일에 변경되었다. 상담 화면이 텍스트 모드가 아니라 아바타 모드로 열리기 시작했고, 아바타 선택지에 권장 표시가 붙었으며, 간단한 이용 안내와 로그인 유도 문구가 추가되었고, 대화 바로 아래에 설문 링크가 배치되었으며, 설문지 앞에 안내 화면이 추가되었다. 검색 설정, 프롬프트, 아바타 정체, 재방문자 로직, 그리고 설문 문항 자체는 변경되지 않았다. 따라서 이 변경은 상담 시스템이 아니라 제시 방식의 변경이지만, 영향이 없는 것은 아니다. 아바타 모드를 기본값으로 두면 아바타를 접한 응답자의 수 자체가 달라지며, 로그인 유도 문구는 로그인하면 아바타가 학생의 이름을 불러준다고 서술하고 있었다. 유지된 응답 가운데 22 건이 6월 10일 이전이고 17 건이 그 이후이며, 그날에는 응답이 기록되지 않았으므로 각 응답을 두 구성 중 하나에 모호함 없이 귀속시킬 수 있다. 5 장은 이 구분이 결과에 관련되는 곳에서 두 기간을 나누어 보고한다.

4.3 분석 절차

분석은 기술적이다. 각 신호에 대해 동의한 응답자의 비율과 그 구간 추정치를 보고하고, 그 비율을 계층별로, 그리고 사전에 고정한 신호 분류에 따라 비교한다. 변수 간 인과 관계는 추정하지 않으며 모형을 적합하지도 않는다.

동의 응답은 동의 수를 해당 문항이 제시된 응답자 수로 나눈 값으로 정의한다. 어떤 응답자에게 해당되지 않았던 문항은 그 문항의 분모에서 제외하며 비동의로 세지 않는다. 분모를 39로 일률 적용하면 학생이 본 적도 없는 문항을 그가 동의하지 않은 신호로 취급하게 되고, 이는 인지와 무관한 이유로 조건부 신호를 과소평가한다. 따라서 분모는 열다섯 개 신호에 대해 39, 두 개에 대해 37, 한 개에 대해 36, 두 개에 대해 6이다.

구간 추정에는 95% 수준의 윌슨 점수법을 사용한다. 응답자가 39 명이고 비율이 척도의 상단에 몰려 있는 경우 정규근사는 구간을 지나치게 좁게 만들고 1을 넘어가기도 한다. 윌슨 구간은 범위가 제한되어 있고 작은 표본과 극단적 비율에서 올바르게 동작하는데, 본 연구의 대다수 신호가 바로 그 상황에 해당한다. 응답 가능자가 6 명인 두 신호의 구간은 그만큼 넓으며, 이 두 점추정치를 나머지와 같은 정밀도를 가진 것처럼 해석하지 않는다.

검정이 아니라 구간을 보고하는 것은, 이 규모의 연구가 가설을 판정하기보다 양을 추정하는 데 쓰이는 편이 낫다는 입장(Anderson and Maxwell, 2016)과, 비율이 특정 경계 위에 있는지에 대한 주장은 기각된 영가설이 아니라 구간이 어디에 놓이는지로 가려야 한다는 입장(Lakens et al., 2018)을 따른 것이다.

설계는 각 신호가 해당되는 사용자의 최소 80%로부터 동의 응답을 받아야 한다는 목표를 두었다. 어떤 신호는 그 구간의 하한이 0.80을 넘을 때에만 목표를 충족한 것으로 센다. 점추정치만이 아니라 구간 전체가 임계값을 넘을 것을 요구하는 것은 보수적인 판독이며, 결과를 제시하기 전에 여기서 고정한다. 계층 값은 해당 계층에 속한 신호들의 동의율 평균으로 보고하므로, 각 신호는 응답자 수와 무관하게 동일한 비중으로 기여한다.

인터페이스의 개별 요소가 전달하는 신호와 시스템 동작의 성질이 전달하는 신호 사이에 동의 응답의 차이가 있는지를 살피기 위해, 스무 개 신호를 동의율을 확인하기 전에 두 집단으로 배정하였다. 사용자가 읽거나 조작할 수 있는 별개의 대상, 즉 라벨·화면·컨트롤·메시지가 전달하는 신호는 요소 기반으로 분류한다. 전용 요소를 갖지 않고 상호작용 전반에 퍼져 있는 성질이며 대개 그것이 나빠질 때에만 드러나는 신호는 동작 기반으로 분류한다. 이 배정으로 각 집단에 열 개씩의 신호가 놓이며 그 목록은 부록에 제시한다. 이 구분은 앞선 배포가 자신의 결과를 설명하기 위해 사용한 바로 그 구분이며 (Authors, 2026), 이를 사전에 고정하는 것이 본 연구로 하여금 그 설명을 되풀이하는 대신 검증할 수 있게 한다. 응답자 39 명은 두 비율 집단 간 차이를 탐지하기에 검정력이 부족하므로, 이 비교는 두 별의 추정치와 그 구간으로 제시하며 유의성 검정을 함께 두지 않는다. 구간을 함께 보고하는 것은 독자가 그것들이 얼마나 겹치는지를 직접 볼 수 있게 하기 위해서이다.

전반 신뢰 문항도 같은 조심스러움으로 보고한다. 단일 폐쇄형 문항은 천장을 부르며, 응답자의 15%를 넘는 인원이 최고 선택지를 고르면 그 도구는 다른 것과의 관계를 볼 해상도를 잃은 것으로 통상 간주된다(Terwee et al., 2007). 본 연구가 그 기준에 대해 어디에 서 있는지는 5.4 절에서 보고한다.

4.2 절에서 하나의 하위집단 분석이 따라 나온다. 수집 도중에 설계가 바뀌었으므로, 신호들은 6 월 10 일 제시 방식 변경 이전과 이후 기간으로 나누어 보고한다. 두 기간 모두 규모가 작으며, 이는 차이를 주장하기 위해서가 아니라 통합 수치의 해석 범위를 한정하기 위해 제시된다.

5. 연구 결과

5.1 표본

유지된 제출 39 건을 분석하였다. 도구는 응답자가 서로 다른 사람인지 확인하지 않았으며, 39 건은 36 개의 서로 다른 브라우저 세션에서 도착하였다. 국적은 자기 기재이며 포함 기준이 아니었다. 설문은 전공 사이트 이용자에게 열려 있었고, 36 명이 국적을 네팔로, 2 명이 한국으로, 1 명이 인도로 기재하였다. 따라서 표본은 하나의 국적 집단에 치우쳐 있으며, 국제학생 일반을 대표하는 것으로 읽어서는 안 된다. 30 명이 경영 트랙을, 7 명이 인공지능 트랙을 지망하였고 2 명은 미정이었다. 22 명이 18 세에서 21 세, 15 명이 22 세에서 25 세였으며 그보다 높은 두 연령대에 각 1 명이 있었다. 남성이 21 명, 여성이 18 명이었다. 완료에는 중위값 2 분 33 초, 평균 2 분 56 초가 걸렸고, 유지된 응답 중 가장 짧은 것이 67 초, 가장 긴 것이 6 분에 조금 못 미쳤다.

39 명 가운데 23 명은 상호작용 모드가 기록되지 않은 상태로 질문에 도달하였으며, 4.1 절에서 밝혔듯 이들은 음성 및 아바타 문항에 기본적으로 응답 가능한 것으로 처리되었다.

5.2 신호별 동의 응답

동의 응답은 스무 개 신호 전반에 걸쳐 높았고 좁은 범위 안에서 변동하였다. 가장 낮은 값은 아바타의 입 움직임과 제스처에 대한 83.3%였고, 가장 높은 값은 각 6 명이 응답한 두 조건부 신호의 100%였다. 스무 개 중 열네 개가 83%에서 95% 사이에 분포하였다. 전반 신뢰는 모든 응답자가 보고하였다. 신호별 동의 응답과 구간은 표 3 에 제시하였고, 같은 값을 두 기준선과 함께 그림 3 에 나타내었다.

4.3 절에서 고정한 판정 규칙을 적용하면 스무 개 중 여섯 개가 설계 목표를 충족하였다. 기관 정체성, 따뜻한 어조, 형식 일관성, 응답 속도 조절, 비회원 열람, 자연스러운 인사이다. 나머지 열네 개는 충족하지 못하였는데, 열네 개 모두 점추정치는 0.80 을 넘지만 구간의 하한이 넘지 못한다. 이는 해당 신호들이 실패했다는 발견이 아니라 응답자 39 명에서 확보되는 정밀도에 관한 진술이다. 이 표본 크기에서 0.90 부근 비율의 구간은 대략 20 퍼센트포인트 폭을 가지므로, 보수적 규칙에서 임계값을 넘으려면 신호가 응답 가능한 거의 모든 사람으로부터 동의 응답을 받아야 한다. 6 명이 응답한 두 조건부 신호가 이 점을 극단적인 형태로 보여준다. 둘 다 응답 가능한 모든 사람으로부터 동의 응답을 받았으나 구간이 61%에서 100%까지 이어지므로, 어느 쪽도 목표를 충족했다고 말할 수 없다.

5.3 요소 기반 신호와 동작 기반 신호

4.3 절에서 고정한 분류는 스무 개 신호를 인터페이스의 개별 요소가 전달하는 열 개와 시스템 동작의 성질이 전달하는 열 개로 나눈다. 요소 기반 신호는 응답 가능자의 평균 93.6%에, 동작 기반 신호는 평균 89.7%로 나타나 예상된 방향으로 3.9 퍼센트포인트의 차이를 보였다. 각 집단 안에서 응답을 통합하면 요소 기반은 92.3%로 구간이 88.9 에서 94.7 까지이고, 동작 기반은 89.8%로 구간이 86.4 에서 92.5 까지이다. 두 구간은 상당 부분 겹치며, 이 표본 크기에서 집단 간 차이를 표집 변동과 구별할 수 없다.

그럼에도 하나의 양상은 일관적이다. 동의 응답이 가장 낮은 세 개 신호가 모두 동작 기반이다. 아바타 립싱크와 제스처 83.3%, 에코 차단 83.8%, 배경 맞춤 답변 84.6%이다. 이어지는 세 신호가 87.2%로 동점이며 그중 둘이 동작 기반, 하나가 요소 기반이다. 요소 기반 신호 중 87.2% 아래로 내려간 것은 없다. 따라서 이 순서는 앞선 국내 배포가 보고한 방향, 즉 상호작용에 분산된 단서가 명시적 라벨로 제시된 단서보다 동의 응답을 덜 받는다는 방향을 재현하되 그 크기는 훨씬 작다. 선행 연구도 동일한 동의형 도구를 사용하였으므로 그 결과를 같은 용어로 다시 서술한다. 앞선 연구는 최고 계층과 최저 계층 사이에 27.5 퍼센트포인트의 격차를 보고하였으나 (Authors, 2026), 여기서 이에 대응하는 폭은 운영 정책 계층 94.2%와 대화 자연스러움 계층 88.3% 사이의 6.0 포인트이다. 우리는 이를 앞선 효과 크기의 확인이 아니라 상당한 감소를 동반한 방향적 재현으로 다룬다.

5.4 계층과 전반 신뢰

계층 평균은 챗봇 정체성 91.5%, 답변 품질 92.9%, 대화 자연스러움 88.3%, 운영 정책 투명성 94.2%, 배경 인식 91.5%였다. 계층 평균은 그림 4에 나타내었다. 따라서 계층들은 서로 가깝게 놓이며, 순서상 대화 자연스러움이 가장 뒤에 온다. 다만 계층 평균이 동작 기반 신호의 비중을 따라가지는 않는다. 동작 기반 비중이 가장 높은 것은 네 개 중 셋인 답변 품질 계층인데 그 계층은 두 번째에 놓인다.

전반 신뢰는 39명 전원이 보고하였다. 이 천장 효과는 방법론적으로 직접적인 결과를 낳는다. 결과 변수에 분산이 없으므로 신호별 동의 응답과 전반 신뢰의 관계를 이 표본에서 추정할 수 없으며, RQ3가 제기한 질문에 공변의 형태로는 답할 수 없다. 말할 수 있는 것은 기술적이지만 나름의 방식으로 더 날카롭다. 모든 응답자가 상담 시스템을 신뢰한다고 보고한 반면, 스무 개 신호 중 열네 개는 설계 목표에 이르지 못했고, 세 개 신호는 해당되는 응답자의 85% 미만으로부터 동의 응답을 받았다. 따라서 이 배포에서 개별 설계 신호가 동의 응답을 받는 것과 호의적인 전반 판단에 이르는 것은 서로 갈라졌다. 어떤 시스템은 그 신뢰를 얻으려고 마련한 구체적 장치 가운데 상당수가 동의를 얻지 못한 채로도 전반적으로 신뢰 받을 수 있다.

5.5 배경 인식 계층

다섯째 계층은 두 수준에서 구현되며, 설문 문항이 주로 첫째 수준을 겨냥하므로 이를 구분해 둘 필요가 있다. 전공 수준에서는 검색 코퍼스 자체가 이 학생 집단을 위해 작성되어 있다. 마흔세 개 청크 가운데 아홉 개가 국제학생을 명시적으로 다루며, 영어 강의, 지역적 진로 이점, 한국에서 공부할 때의 실질적 관심사를 포괄한다. 이러한 주제를 묻는 학생은 계정 보유 여부와 무관하게 자신의 처지에 맞추어 작성된 답변을 받는다. 개인 수준에서는 가입 시 기록된 국가와 관심 트랙이 답변 구성 시 활용되며, 재방문 인사는 서로 다른 방문일이 두 번째가 되는 시점부터 표시된다.

측정된 동의 응답은 배경 인지가 89.7%, 배경 맞춤 답변이 84.6%였다. 배경 인지 89.7%는 조건부가 아닌 나머지 열세 개 신호의 범위인 87.2~97.4% 안에 있으나, 배경 맞춤 답변 84.6%는 그 열세 개 어느 것보다도 낮아 비조건부 신호 가운데 최저값이다. 둘 다 보수적 규칙에서 설계 목표를 넘지는 못한다. 재방문자 인사는 응답 가능한 여섯 명 전원으로부터 동의 응답을 받았고, 관측 여섯 개가 합의하는 넓은 구간을 가진다.

이 두 문항은 상담 시스템이 응답자의 배경을 알고 있는 것처럼 보였는지, 그리고 답변이 자신 같은 사람을 향한 것으로 느껴졌는지를 묻는다. 두 가지 모두 계정 수준의 개인화만이 아니라 전공 수준의 작성 방식으로도 충족될 수 있다. 따라서 위 수치는 계정 수준 기제에 관한 근거라기보다 이 계층 전체에 관한 근거로 읽어야 한다. 둘을 분리하려면 무엇이 필요한지는 6.4 절에서 다시 다룬다.

5.6 6월 10일 제시 방식 변경

22 건의 응답이 제시 방식 변경 이전이고 17 건이 그 이후이다. 대부분의 신호는 두 기간 사이에서 거의 움직이지 않는다. 세 개가 하락하였고 그중 둘은 15 퍼센트포인트를 넘게 움직였다. 매끄러운 모드 전환이 95.5%에서 76.5%로, 개인정보 동의 화면이 100%에서 82.4%로 내려갔다. 자연스러운 인사는 100%에서 88.2%로 11.8 퍼센트포인트 하락하였다. 이 세 인터페이스 요소 자체는 6월 10일에 변경되지 않았다.

타당한 해석은 이 변경이 각 부분의 동작 방식이 아니라 응답자가 인터페이스의 어느 부분을 접했는지를 바꾸었다는 것이다. 상담 화면이 텍스트 모드가 아니라 아바타 모드로 열리고 아바타에 권장 표시가 붙으면, 응답자가 모드 전환을 실제로 사용해 보았을 가능성이 낮아진다. 모드 전환 문항이 묻는 것이 바로 그것이다. 또한 대화 중에 추가된 설문 링크는 대화에서 설문지에 이르는 경로를 짧게 만들므로, 후반 기간의 응답자는 동의 고지와 인사말이 놓인 가입 및 진입 화면을 거치지 않고 설문문에 도달했을 수 있다. 두 해석 모두 사후적이고 하위집단은 작으며 구간은 넓다. 우리가 이 차이를 보고하는 것은 이 비교가 어떤 효과를 식별하기 때문이 아니라, 수집 도중에 설계가 바뀌었으므로 통합 수치만 제시하는 것이 오토크를 낳기 때문이다.

5.7 자유응답

15 명이 긍정 문항에, 14 명이 부정 문항에 무언가를 적었다. 상당수는 서술이라기보다 짧은 인사에 가깝고, 부정 문항의 여러 응답은 신뢰가 가지 않은 지점이 없었다고 적고 있으므로, 이 자료는 분석보다 예시를 뒷받침한다.

내용이 있는 긍정 응답 가운데 응답자들이 가장 자주 언급한 것은 답변의 명료함과 유용함이었으며, 두 명은 상담 시스템이 모르는 것을 다루는 방식을 언급하였다. 한 응답자는 이 시스템이 여러 종류의 정보를 제공하며 상담 시스템이 무언가를 모를 때에는 이용자를 안내해 준다고 적었는데, 이는 한계 인정 신호를 의도한 대로 기술한 것이다. 내용이 있는 부정 응답 가운데 응답 지연이 두 번 언급되었다. 두 응답은 코퍼스의 경계를 이용자 쪽에서 기술한다. 한 응답자는 질문이 상담 시스템이 아는 범위를 벗어났을 때 직접적인 답을 주지 않았다고 보고하였고, 다른 응답자는 답변이 때때로 반복적이거나 구체적인 세부가 부족했다고 적었다. 한 응답은 음성이 다소 컸다고 지적하였다. 이들은 3.2 절에서 기술한 검색 임계값 동작, 그리고 속도 조절 및 음성 출력 신호에 대응하며, 동일한 장치가 어떤 응답자에게는 정직함으로, 다른 응답자에게는 회피로 경험될 수 있음을 보여준다.

6. 논의 및 결론

6.1 결과가 보여주는 것

아래 결과를 읽는 방식에 두 가지 약속이 놓여 있다. 각 신호는 계층 안에서, 각 계층은 설계 안에서 같은 무게를 갖는다. 데이터를 보기 전에 어떤 단서도 더 진단적이라고 취급하지 않는다는 뜻이다. 예측 변수의 상대적 중요도가 미리 확립되지 않은 상황에서 동일 가중치는 방어 가능한 기본값이며, 작은 표본에서 추정된 가중치보다 나은 경우가 많다(Dawes, 1979). 그리고 동의 응답 값은 단일 지수가 아니라 신호 단위로 보고한다. 총합은 6.3 절이 딛고 선 그 편차를 가려 버리기 때문이다.

이제 세 연구 질문에 답할 수 있다. RQ1 은 배경 인식 계층을 더한 신뢰 지향 아바타 상담봇을 국제 대학 생용으로 설계하고 배포할 수 있는지를 물었다. 3 장은 그러한 시스템을 보고한다. 검색과 생성을 기관이 관리하는 하드웨어에서 수행하는 네 계층, 하나의 백엔드 위에서 동작하는 세 가지 상호작용 모드, 그리고 구현 위치와 함께 계층별로 규정된 스무 개의 신호이다. 이 시스템은 9 일 동안 실제 학생을 대상으로 운영되었고 여기서 분석한 기록을 산출하였다. 그런 의미에서 RQ1 은 긍정적으로 답해지되, 구축이 효과의 입증과 같지 않다는 단서가 붙는다.

RQ2 는 각 신호가 사용자에게 얼마나 동의 응답을 받았는지, 그리고 개별 요소로 제시된 신호와 시스템 동작이 전달하는 신호 사이에 체계적인 차이가 있는지를 물었다. 동의 응답은 83.3%에서 100% 사이로 높고 좁게 분포하였다. 스무 개 중 여섯 개가 사전에 고정한 보수적 규칙에서 설계 목표를 충족하였으며, 나머지 열네 개가 충족하지 못한 것은 동의 응답이 낮아서라기보다 이 표본 크기에서의 구간 폭에 기인한다. 두 번째 부분에 대해서는, 요소 기반과 동작 기반 신호의 차이가 3.9 퍼센트포인트로 예상된 방향에 있고, 동의 응답이 가장 낮은 세 신호가 모두 동작 기반이고 이어지는 세 신호가 87.2%로 동점이며, 두 집단의 신뢰구간은 겹친다. 앞선 배포가 보고한 양상 (Authors, 2026)은 방향에서 재현되고 크기에서 크게 감쇠되었다.

RQ3 은 신호별 동의 응답이 전반 신뢰와 어떻게 연결되는지, 그리고 배경 인식 신호가 얼마나 동의 응답을 받았는지를 물었다. 배경 신호는 각각 응답자의 89.7%와 84.6%로부터 동의 응답을 받았고 재방문자 인사는 응답 가능한 여섯 명 전원으로부터 받았으므로, 이 계층은 그것이 더해진 네 계층과 같은 구간에서 작동하였다. 앞부분은 설정한 형태로는 답할 수 없다. 전반 신뢰를 모든 응답자가 보고하여 결과 변수에 분산이 없으므로 공변을 추정할 수 없다. 말할 수 있는 것은 둘이 함께 움직이지 않았다는 점이며, 이는 6.3 절에서 다룬다.

6.2 구현과 동의 응답

본 연구는 두 가지 물음으로 구성되며, 결과는 두 물음 모두에 직접 답한다. 첫째는 신뢰 문헌을 개별 발견들의 집합으로 두지 않고 하나의 시스템으로 옮길 수 있는가이다. 다섯 계층에 걸쳐 스무 개의 신호를 규정하였고, 각 신호는 계층 수준의 근거와 명시된 구현 위치를 가지며, 스무 개 모두가 구축되어 배포되었다. 3.4 절이 이를 열거하고, 부록이 각각을 측정하는 설문 문항을 담는다. 여기서 기여는 그 개수에 있

다. 하나의 기제를 따로 떼어 평가한 것이 아니라, 실제 학생을 대상으로 돌아가는 하나의 시스템이 스무 개를 동시에 실행 나른 것이다.

둘째는 그 신호들이 대상에게서 동의 응답을 받았는가이다. 받았고, 고르게 받았다. 모든 신호가 해당되는 응답자의 83.3%에서 100% 사이에서 보고되었고, 다섯 계층은 88.3%에서 94.2% 사이에 놓이며, 구현이 표면화되지 못했다고 판단할 만한 구간으로 떨어진 신호는 없었다. 스무 개의 신호를 동시에 실행하는 시스템에서 이것이 실질적인 결과이다. 어떤 신호가 인지되기 위해 다른 신호를 희생해야 하지 않았다.

다만 이 동의 응답 수치는 보이는 것보다 여유가 적다. 4.3 절에서 고정한 규칙에서는 스무 개 중 여섯 개가 0.80 설계 목표를 넘는데, 5.2 절에서 밝힌 구간 폭 때문이다. 이 규모의 연구는 동의 응답이 높고 고르다는 것을 확립할 수 있으나, 개별 신호를 임계값에 대해 입증하지는 못한다. 이는 정밀도에 관한 진술이며, 해결책은 설계 변경이 아니라 표본 크기이다.

6.3 동의 응답과 신뢰는 별개의 결과이다

전반 신뢰는 모든 응답자가 보고한 반면, 스무 개 신호 중 열네 개는 설계 목표를 넘지 못했고 가장 낮은 신호는 대략 여섯 명 중 한 명이 동의하지 않았다. 따라서 두 측정치는 함께 움직이지 않았다. 전반 판단은 균일했던 반면, 그것을 얻기 위해 구축한 개별 장치에 대한 동의는 그렇지 않았다.

두 가지 해석이 가능하며 본 연구는 이를 분리하지 못한다. 신뢰가 거의 모든 사람이 동의한 소수의 신호에 놓여 있을 수 있으며, 목표를 넘은 여섯 개 가운데 둘이 답변 품질 계층에 속하고 둘이 운영 정책 계층에 속하므로, 답변 품질 계층은 후보의 하나가 된다. 또는 세션 직후에 묻는 단일 이분 문항이 수업에서 방금 소개받은 시스템에 대한 만족에 가까운 무언가를 재고 있을 수 있으며, 그렇다면 그 천장은 설계가 아니라 상황을 반영한다. 어느 해석에서든 실무적 함의는 같다. 총합 신뢰 점수는 개별 설계 신호가 작동한다는 근거가 아니며, 본 연구가 그렇게 한 것처럼 둘을 나누어 보고해야 한다.

세 번째 가능성이 두 해석 모두에 걸린다. 동의형 문항은 집단주의와 권력거리가 높은 문화권 응답자가 더 쉽게 동의하는 쪽으로 답하며, 제 2 언어로 시행할 경우 응답이 같은 방향으로 움직인다(Harzing, 2006). 본 표본은 거의 전부가 그러한 환경에서 왔고 영어로 응답하였으므로, 전반 신뢰 문항의 균일함을 설계만으로 돌릴 수 없다.

6.4 연구의 한계

표본이 작고 좁다. 응답자 39 명이 낳는 구간 폭은 5.2 절에서 밝힌 대로이며, 이 때문에 대부분의 신호를 0.80 목표에 대해 어느 방향으로도 판정할 수 없다. 39 명 중 36 명이 동일한 국적을 기재하였으므로 이 결과는 하나의 프로그램에 속한 하나의 집단을 기술하는 것이지 국제학생 일반을 기술하지 않는다. 참여는 자발적이었고 수업 세션을 중심으로 이루어졌으므로, 이미 시스템에 호의적인 학생이 유리하게 포함된다.

도구에는 세 가지 구체적인 약점이 있다. 4.1 절에서 밝혔듯 지각이 아니라 동의를 기록하므로, 인지되었으나 부정적으로 판단된 신호와 인지되지 않은 신호가 같은 방식으로 기록된다. 이분형이므로 정도를 기록하지 못하고 희미하게 인지된 신호와 분명하게 인지된 신호를 구별하지 못한다. 그리고 단일 전반 신뢰 문항이 천장에 있어 RQ3가 요구한 결과 변수의 분산이 사라졌다.

배포는 추론을 더 제약한다. 설문을 여는 데 최소 대화량이 필요하지 않았으므로, 일부 응답자는 거의 사용하지 않은 시스템에 대해 답했을 수 있다. 도구에는 동일 브라우저의 두 번째 제출을 막는 장치가 없었고 39 건의 응답이 36 개 브라우저 세션에서 도착하였다. 음성 및 아바타 문항의 응답 가능 여부 기록은 브라우저에 보관되었고 기록이 없으면 응답 가능으로 기본 처리되었으므로, 세 신호의 분모에 포함된 응답자 가운데 23명은 해당 모드를 사용한 기록이 없다. 이 세 신호를 해당 모드 사용 기록이 있는 응답자로 한정할 완전사례 점검을 수행하였다. 동의 응답은 양방향으로 움직였다. 에코 차단은 83.8%에서 78.6%로, 아바타 립싱크와 제스처는 83.3%에서 84.6%로, 발음 정규화는 89.2%에서 100%로 바뀌었다. 어느 규칙에서도 설계 목표에 대한 판정이 달라진 신호는 없다. 완전사례의 분모는 13 또는 14 이므로 그 구간은 5.2 절에서 보고한 것보다 넓다. 이 점검은 fallback 규칙을 대체하는 것이 아니라 그 영향의 범위를 한정한다. 시스템의 제시 방식이 수집 도중에 변경되었으며, 검색·프롬프트·아바타 구성은 손대지 않았으나 기본 모드와 설문 진입 경로는 그렇지 않았다.

설계 자체에서 비롯되는 제약이 하나 더 있다. 계정 없이 열람하는 것은 스무 개 신호 가운데 하나(F14)이며 응답자의 94.9%로부터 동의 응답을 받아 설계 목표를 충족한 여섯 개에 든다. 따라서 배포는 로그인을 요구하지 않았고, 39 명 가운데 6 명이 응답 시점에 계정을 보유하고 있었다. 나머지 33 명에게는 가입 시 기록된 국가와 관심 트랙을 답변으로 옮기는 배경 인식 계층의 계정 수준 기제가 작동하지 않았으나, 5.5 절에서 기술한 전공 수준 기제는 작동하였다. 검색 코퍼스 43 개 청크 가운데 9 개가 국제학생을 위해 작성되었고 계정 상태와 무관하게 반환되기 때문이다. 그러므로 두 개의 배경 문항은 두 기제 가운데 하나만 작동할 수 있었던 응답자가 다수인 상태에서 응답되었다. 해당 응답은 기록된 그대로 보고하며 어느 기제에 귀속시키지 않는다. 두 기제를 분리하는 데 필요한 기록은 6.5 절에 제시한다.

인터페이스에도 같은 종류의 제약이 있다. 모드 전환은 신호 F12이며 87.2%로 설계 목표에 미치지 못하였다. 세 모드는 하나의 검색·생성 경로를 공유하므로 답변의 사실 내용은 선택한 모드에 좌우되지 않으나, 대화 상태는 공유하지 않는다. 대면형에서 음성 또는 텍스트로 옮긴 학생은 앞선 대화 없이 새 모드를 시작한다. 이는 설계의 성질이 아니라 배포 시점의 구현이 달을 수 있던 선이었으며, 여기에 기술적 한계로 기록하고 6.5 절에 개발 과제로 넘긴다.

표본의 좁음은 두 방향으로 작동한다. 거의 전부가 네팔에서 온 집단은 국제학생 일반에 관한 주장의 근거로는 빈약하다. 그러나 동시에 이 문헌이 거의 닿지 못한 모집단이기도 하다. 행동과학, 특히 인간-컴퓨터 상호작용의 표본은 압도적으로 서구·고학력·산업화·부유·민주주의 환경에서 추출되어 왔고, 그 위에 세운 일반화는 다른 곳에서 반복적으로 성립하지 않았다(Henrich et al., 2010; Linxen et al., 2021). 이 집단을 직접 다룬 유일한 비교 연구는 네팔과 영국 응답자에게서 생성형 도우미 수용을 이끄는 요인이 서로 달랐음을, 곧 동일한 기술수용 구성개념이 두 나라에서 다른 무게를 가짐을 보고했다

(Budhathoki et al., 2024). 단일 국가 표본으로는 드러낼 수 없는 종류의 차이이다. 여기서 보고하는 것은 통상적인 의미에서 좁으면서 구성에서는 드문 것이며, 둘 중 하나가 다른 하나를 대신하지 않도록 함께 진술해야 한다. 실증 논문은 결과가 어느 참가자·재료·맥락에 적용되도록 의도되었는지를 밝혀야 한다는 권고(Simons et al., 2017)에 따라, 본 연구의 발견은 자신들을 위해 작성된 검색 코퍼스를 갖춘 전공에서 제 2 언어로 상담 도우미를 사용하는 남아시아·동아시아 학부생에게 적용되는 것으로 제시한다.

끝으로 본 연구는 전체적으로 기술적이다. 어떤 신호가 신뢰에 미치는 효과에 대해서도 인과적 주장을 하지 않으며, 계층 구조는 검증된 측정 모형이 아니라 설계 도식이다. 요인분석은 수행하지 않았고 이 표본 크기에서는 정당화되지도 않는다.

6.5 후속 연구

가장 분명한 다음 단계는 표본 크기이다. 200 명에서 300 명 규모의 배포라면 구간이 충분히 좁아져 개별 신호를 입증할 수 있고, 여기서 3.9 퍼센트포인트인 요소 기반과 동작 기반의 차이도 검증할 수 있다. 이 차이를 안정적으로 탐지하려면 표본이 몇 배는 커야 한다.

도구도 보강해야 한다. 문항은 정도를 복원할 수 있도록 등급 응답과 짝지어야 하며, 일부를 역코딩하여 무차별적 응답을 직접 탐지할 수 있게 해야 한다. 전반 신뢰는 단일 이분 문항이 아니라 다문항 척도로 측정해야 하고, 이는 본 연구에 없었던 결과 변수의 분산을 회복시킨다.

배경 인식 계층에 대해서는 5.5 절에서 기술한 전공 수준 기제와 계정 수준 기제를 설계상 분리해야 한다. 로그인을 필수로 하거나, 계정 맥락이 실제로 전송되었는지를 세션 단위로 기록하면, 계층 전체가 아니라 각 기제에 대해 동의 응답을 보고할 수 있다. 동일한 기록은 조건부 신호에도 도움이 된다. 본 연구에서 그 응답 가능 여부는 브라우저에 보관된 기록으로부터 추정되었다.

모드 간 연속성도 해결해야 한다. 모드를 전환할 때 대화 상태를 이어받아 대면형에서 텍스트로 옮긴 학생이 다시 시작하지 않고 이어가게 하면 6.4 절에 기록한 단절이 사라지며, 이는 모드 선택 자체가 동의 응답에 영향을 주는지를 묻기 위한 전제 조건이기도 하다.

두 기제를 분리하면 배경 인식 계층을, 에이전트가 무엇을 아는지가 아니라 누구로 보이는지를 조작하는 연구와 견줄 수도 있다. 아바타의 정체성이 인식적 신뢰와 답변 수용을 움직인다는 것이 보고된 바 있다(Anthis and Kyriakidou-Zacharoudiou, 2026). 신호 단위로 잰 동의 응답과 정체성으로 조작한 신뢰는 같은 질문을 잡는 서로 보완적인 손잡이이며, 둘을 함께 실은 설계는 어느 한쪽만으로 말할 수 있는 것보다 많은 것을 말할 것이다.

측정을 넘어서, 재현은 확장되어야 한다. 요소 기반과 동작 기반 차이의 방향은 이제 코호트·언어·도메인이 다른 두 배포에서 관찰되었으나 그 크기는 상당히 달랐다. 몇 퍼센트포인트의 차이를 탐지할 검정력을 갖춘 세 번째 배포가 어느 크기가 안정적인 것인지를 확정할 것이다.

6.6 결론

본 연구는 두 가지를 하려 하였고 그 둘 모두를 보고한다. 하나는 개별 기제별로 조직된 문헌을 하나의 작동하는 시스템으로 옮긴 것이다. 다섯 계층에 걸친 스무 개의 신뢰 신호를 각각의 근거와 구현 위치와 함께 규정하였고, 전공이 학생에 대해 아는 것을 답변으로 옮기는 배경 인식 계층을 더하였다. 다른 하나는 구축한 것이 그 대상에게서 동의 응답을 받았는지를 신호 하나하나에 대해 측정한 것이다.

그러하였다. 스무 개 신호 각각이 해당되는 응답자의 83.3%에서 100% 사이에서 보고되었고, 다섯 계층은 서로 6 퍼센트포인트 안에 놓인다. 동의 응답이 눈에 띄는 신호에 몰리면서 나머지를 희생시키지 않았다. 이는 계층적 설계가 낳고자 하는 결과이며, 이만큼 많은 신호를 동시에 싣는 시스템에서 이전에 입증된 바 없는 결과이다.

이 결과에는 두 가지 단서가 따른다. 응답자 39 명은 구간을 넓게 만들어 여섯 개 신호만이 0.80 설계 목표에 대해 인증될 수 있으므로, 발견은 동의 응답이 높고 고르다는 것이지 각 신호가 개별적으로 기준을 충족한다는 것이 아니다. 또한 개별 신호가 편차를 보인 반면 전반 신뢰는 모든 응답자가 보고하였으므로, 총합 판단을 부분들이 작동한다는 근거로 사용할 수 없다. 본 논문이 그렇게 한 것처럼 둘을 나누어 보고하는 것이 어느 쪽이든 정직하게 읽을 수 있게 한다.

표 1. 계층별 근거

계층	이 계층이 다루는 것	근거
L1 챗봇 정체성	학생이 누구와 이야기하는가, 그것이 고지되는가	임베디먼트된 인간형 표상이 사회적 실재감을 높이고, 이것이 신뢰 믿음을 높인다 (Qiu & Benbasat, 2009; Li et al., 2023)
L2 답변 품질	답변이 기관 자료에 근거하는가, 근거가 없을 때 한계를 두는가	검색된 문서에 조건화한 생성이 더 사실적이고 환각이 적다 (Lewis et al., 2020; Shuster et al., 2021). 자동화의 역량 한계를 전달하는 것이 최대가 아닌 적정 수준의 신뢰를 지지한다 (Lee & See, 2004)
L3 대화 자연스러움	주고받음이 대화처럼 작동하는가	이미지 없음에서 정지 초상, 애니메이션 아바타로 수준을 올리면 사회적 실재감이 높아진다 (Li et al., 2023)
L4 운영 정책 투명성	학생의 데이터와 접근을 시스템이 어떻게 다루는가	프라이버시 우려는 대화형 에이전트 신뢰의 부적 결정요인이다 (Li et al., 2023). 관계적 대화 행동이 반복 세션에 걸친 관계를 유지시킨다 (Bickmore & Picard, 2005). 신뢰는 교육용 AI에 대한 행동의도를 예측한다 (Ayanwale & Molefi, 2024)
L5 배경 인식	시스템이 학생에 대해 아는 것을 세션을 넘어 이어가는가	확립된 선례 없음. 선행 개인화는 시스템이 이용 중에 측정된 상태에 적응한다 (Tielman et al., 2019; Verhagen et al., 2023)

표 2. 20 개 신호와 구현 위치

ID	계층	신호	구현 위치
F1	L1	전신 디지털 트윈	영상 영역에 렌더링되는 사진 기반 아바타 세션
F2	L1	기관 정체성	상단 바와 페이지 구성에 표시된 프로그램 정체성
F3	L1	AI 고지	시스템 프롬프트와 인터페이스에서 AI임을 명시
F4	L2	검색 그라운드링	코사인 0.40 이상 상위 3개 청크를 프롬프트에 주입
F5	L2	한계 인정	임계값 미달 시 출처를 주입하지 않고 정보가 없다고 답함
F6	L2	따뜻한 어조	시스템 프롬프트에 어조 지정
F7	L2	형식 일관성	시스템 프롬프트에 고정된 답변 구조
F8	L3	응답 속도 조절	스트리밍 방식 응답 전달
F9	L3	다중 예코 차단	아바타 발화 중 음성 입력 억제
F10	L3	발화 중단 제어	아바타 발화를 도중에 멈추는 컨트롤
F11	L3	아바타 립싱크·제스처	발화와 동기화된 아바타 영상
F12	L3	매끄러운 모드 전환	하나의 백엔드 위 FTF·STS·TTT 모드
F13	L4	개인정보 동의 UI	가입 시점에 동의 제시
F14	L4	비회원 열람	계정 없이 프로그램 페이지

			이용 가능
F15	L4	자연스러운 인사	진입 시 인사 문구
F16	L4	방문 횟수 기록	계정별 방문일 수 집계
F17	L4	TTS 발음 정규화	음성 합성 전 텍스트 정규화
F18	L5	배경 인지	가입 시 국가·관심 트랙을 답변 생성 시 활용
F19	L5	배경 맞춤 답변	저장된 배경에 맞춘 답변
F20	L5	재방문자 인식	두 번째 방문일부터 인사 표시

표 3. 신호별 동의율과 95% 신뢰구간 (N = 39)

ID	계층	신호	분류	동의율	95% CI	분모	목표
F1	L1	전신 디지털 트윈	동작	87.2%	[73.3, 94.4]	39	
F2	L1	기관 정체성	요소	97.4%	[86.8, 99.5]	39	O
F3	L1	AI 고지	요소	89.7%	[76.4, 95.9]	39	
F4	L2	검색 그라운드	동작	87.2%	[73.3, 94.4]	39	
F5	L2	한계 인정	요소	89.7%	[76.4, 95.9]	39	
F6	L2	따뜻한 어조	동작	97.4%	[86.8, 99.5]	39	O
F7	L2	형식 일관성	동작	97.4%	[86.8, 99.5]	39	O
F8	L3	응답 속도 조절	동작	97.4%	[86.8, 99.5]	39	O
F9	L3	다층 에코 차단	동작	83.8%	[68.9, 92.3]	37	
F10	L3	발화 중단 제어	요소	89.7%	[76.4, 95.9]	39	
F11	L3	아바타 립싱크·제스처	동작	83.3%	[68.1, 92.1]	36	
F12	L3	매끄러운 모드 전환	요소	87.2%	[73.3, 94.4]	39	
F13	L4	개인정보 동의 UI	요소	92.3%	[79.7, 97.3]	39	
F14	L4	비회원 열람	요소	94.9%	[83.1, 98.6]	39	O
F15	L4	자연스러운 인사	요소	94.9%	[83.1, 98.6]	39	O
F16	L4	방문 횟수 기록	요소	100.0%	[61.0, 100.0]	6	
F17	L4	TTS 발음 정교화	동작	89.2%	[75.3, 95.7]	37	
F18	L5	배경 인지	동작	89.7%	[76.4, 95.9]	39	
F19	L5	배경 맞춤 답변	동작	84.6%	[70.3, 92.8]	39	
F20	L5	재방문자 인식	요소	100.0%	[61.0, 100.0]	6	
F21	-	전반 신뢰	-	100.0%	[91.0, 100.0]	39	

* 목표 = Wilson 95% CI 하한이 0.80 을 넘는 경우에만 O (4.3 절 판정 규칙). 분류: 요소 = 요소 기반, 동작 = 동작 기반.

부록. 설문 문항 전문

「기록 대상」 열은 각 문항이 응답자에게 무엇을 확인하도록 요구하는지를 밝힌 것이다. 신호의 존재나 인지를 직접 묻는 문항은 다섯 개이며, 나머지는 평가·선호·효과를 묻는다.

응답 형식은 Agree / Disagree 두 버튼(중간값·무응답 없음). 인구통계 5 문항(국가·트랙·연령대·성별·MBTI)은 별도.

설문 문항 (제시된 원문)	ID	기록 대상
Talking with this assistant felt like talking to the program's real avatar host.	F1	평가
This assistant felt like an official tool of [university name].	F2	평가
The assistant made it clear that it is an AI, not a real person.	F3	존재·인지
The assistant's answers felt accurate and fact-based.	F4	평가
When it did not know something, the assistant honestly said so or suggested contacting the program staff.	F5	존재·인지
The assistant's tone felt warm and friendly.	F6	평가
The assistant's answers were consistent and professional in style.	F7	평가
The pacing of the conversation (time from my question to its reply) felt natural.	F8	평가
During voice conversation, my voice and the assistant's voice never overlapped or cut off awkwardly. [조건부]	F9	존재·인지
When I wanted to interrupt while the assistant was speaking, I was able to (or felt I could).	F10	존재
The avatar's lip movements and expressions looked natural. [조건부]	F11	평가
I liked being able to switch freely between Avatar, Voice, and Text modes.	F12	선호
At sign-up, the data-handling notice (what is collected, why, and for how long) was presented clearly.	F13	평가
Being able to browse without logging in made me feel more at ease.	F14	효과
I liked that the assistant greeted me with a natural, friendly welcome message.	F15	선호
It felt friendly that the assistant remembered how many times I had visited. [조건부]	F16	효과
The assistant read out abbreviations (AI, GPT, etc.) naturally when speaking. [조건부]	F17	평가
The assistant seemed aware of my background and interests.	F18	존재·인지
The answers felt tailored to someone like me, rather than generic.	F19	평가

I felt the site recognized me as a returning or known visitor. [조건부]	F20	존재·인지
Overall, I felt I could trust this assistant.	F21	전반
What moment or feature felt most trustworthy? [자유응답]	-	
What felt least trustworthy or awkward? [자유응답]	-	

참고문헌

36 편 전부 Crossref·OpenAlex·Semantic Scholar 등 복수 색인에 실접속하여 검증 (2026-08-12). * = 프리프린트(동료심사 전).

- Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, 21(1), 1–12. <https://doi.org/10.1037/met0000051>
- Anthis, Z., & Kyriakidou-Zacharoudiou, A. (2026). Face value: How avatar identity shapes epistemic trust in AI-mediated learning. *Computers and Education: Artificial Intelligence*, 10, Article 100610. <https://doi.org/10.1016/j.caeai.2026.100610>
- Ayanwale, M. A., & Molefi, R. R. (2024). Exploring intention of undergraduate students to embrace chatbots: From the vantage point of Lesotho. *International Journal of Educational Technology in Higher Education*, 21(1), Article 20. <https://doi.org/10.1186/s41239-024-00451-8>
- Béchar, P., & Marquez Ayala, O. (2024). Reducing hallucination in structured outputs via Retrieval-Augmented Generation. In Y. Yang, A. Davani, A. Sil, & A. Kumar (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)* (pp. 228–238). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.naacl-industry.19>
- Bickmore, T. W., & Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction*, 12(2), 293–327. <https://doi.org/10.1145/1067860.1067867>
- Bilquise, G., Ibrahim, S., & Salhieh, S. M. (2024). Investigating student acceptance of an academic advising chatbot in higher education institutions. *Education and Information Technologies*, 29(5), 6357–6382. <https://doi.org/10.1007/s10639-023-12076-x>
- Budhathoki, T., Zirar, A., Njoya, E. T., & Timsina, A. (2024). ChatGPT adoption and anxiety: A cross-country analysis utilising the unified theory of acceptance and use of technology (UTAUT). *Studies in Higher Education*, 49(5), 831–846. <https://doi.org/10.1080/03075079.2024.2333937>
- Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34(7), 571–582. <https://doi.org/10.1037/0003-066X.34.7.571>
- Fogg, B. J. (2003). Prominence-interpretation theory: Explaining how people assess credibility online. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems* (pp. 722–723). ACM Press. <https://doi.org/10.1145/765891.765951>
- George, J. F., Mirsadikov, A., & Mennecke, B. E. (2016). Website credibility assessment: An empirical investigation of prominence-interpretation theory. *AIS Transactions on Human-Computer Interaction*, 8(2), 40–57. <https://doi.org/10.17705/1thci.00078>
- Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2), 627–660. <https://doi.org/10.5465/annals.2018.0057>
- Harzing, A. W. (2006). Response styles in cross-national survey research. *International Journal of Cross Cultural Management*, 6(2), 243–266. <https://doi.org/10.1177/1470595806066332>
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2–3), 61–83. <https://doi.org/10.1017/S0140525X0999152X>

- Hiratsuka, H. (2016). Japanese language as an organizational barrier for international students to access to university services: A case of Aoyama Gakuin University. *Journal of International Students*, 6(4), 956–966. <https://doi.org/10.32674/jis.v6i4.328>
- Hornbæk, K., Sander, S. S., Bargas-Avila, J. A., & Simonsen, J. G. (2014). Is once enough? On the extent and content of replications in human-computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)* (pp. 3523–3532). ACM. <https://doi.org/10.1145/2556288.2557004>
- Kalashi, K., Montazer, G. A., Saed, S., & Olyaei, S. (2025). Adoption of chatbots in higher education: Global insights and recommendations for Iran. In *2025 12th International and the 18th National Conference on E-Learning and E-Teaching (ICeLeT)* (pp. 1–6). IEEE. <https://doi.org/10.1109/ICeLeT66022.2025.11025265>
- Kim, H., Han, B., Kim, J., Syawaludin, M. F., Kim, G. J., & Hwang, J.-I. (2024). Engaged and affective virtual agents: Their impact on social presence, trustworthiness, and decision-making in the group discussion. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)* (pp. 1–17). ACM, New York, Article 149. <https://doi.org/10.1145/3613904.3642917>
- Kirlappos, I., Sasse, M. A., & Harvey, N. (2012). Why trust seals don't work: A study of user perceptions and behavior. In S. Katzenbeisser, E. Weippl, L. J. Camp, M. Volkamer, M. Reiter, & X. Zhang (Eds.), *Trust and Trustworthy Computing (TRUST 2012)*, *Lecture Notes in Computer Science*, vol. 7344 (pp. 308–324). Springer. https://doi.org/10.1007/978-3-642-30921-2_18
- Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2), 259–269. <https://doi.org/10.1177/2515245918770963>
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems*, vol. 33 (pp. 9459–9474). Curran Associates.
- Li, Q., Luximon, Y., & Zhang, J. (2023). The influence of anthropomorphic cues on patients' perceived anthropomorphism, social presence, trust building, and acceptance of health care conversational agents: Within-subject web-based experiment. *Journal of Medical Internet Research*, 25, Article e44479. <https://doi.org/10.2196/44479>
- Liew, T. W., Tan, S.-M., & Ismail, H. (2017). Exploring the effects of a non-interactive talking avatar on social presence, credibility, trust, and patronage intention in an e-commerce website. *Human-centric Computing and Information Sciences*, 7(1), Article 42. <https://doi.org/10.1186/s13673-017-0123-4>
- Linxen, S., Sturm, C., Brühlmann, F., Cassau, V., Opwis, K., & Reinecke, K. (2021). How WEIRD is CHI? *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14. <https://doi.org/10.1145/3411764.3445488>
- Lukyanenko, R., Maass, W., & Storey, V. C. (2022). Trust in artificial intelligence: From a foundational trust framework to emerging research opportunities. *Electronic Markets*, 32(4), 1993–2020. <https://doi.org/10.1007/s12525-022-00605-4>

- Marconi, L., Longo, L., & Cabitza, F. (2026). Assessing interaction quality in human-AI dialogue: An integrative review and multi-layer framework for conversational agents. *Machine Learning and Knowledge Extraction*, 8(2), Article 28. <https://doi.org/10.3390/make8020028>
- Newton, D. C., Tomin, A. J., & LaMontagne, A. D. (2021). Exploring the challenges and opportunities for improving the health and wellbeing of international students: Perspectives of international students. *Journal of the Australian and New Zealand Student Services Association*, 29(1), 18-34. <https://doi.org/10.30688/janzssa.2021.1.02>
- Ng, S. W. T., & Zhang, R. (2025). Trust in AI chatbots: A systematic review. *Telematics and Informatics*, 97, Article 102240. <https://doi.org/10.1016/j.tele.2025.102240>
- Nourian, L., Callis, A., Patterson, S., Miao, J., Heard, J., & Tigwell, G. W. (2026). Understanding how international students in the U.S. are using conversational AI to support cross-cultural adaptation. *arXiv:2605.15127*. <https://doi.org/10.48550/arXiv.2605.15127> (preprint).
- Pereira, D. S. M., Falcão, F., Costa, L., Lunn, B. S., Pêgo, J. M., & Costa, P. (2023). Here's to the future: Conversational agents in higher education — a scoping review. *International Journal of Educational Research*, 122, Article 102233. <https://doi.org/10.1016/j.ijer.2023.102233>
- Perry, C. J., Lausch, D., McKim, C. A., & Weatherford, J. (2020). Knowledge, use, and perceived value of university student services. *Journal of International Students*, 10(3), 613-628. <https://doi.org/10.32674/jis.v9i4.269>
- Polyportis, A., & Pahos, N. (2025). Understanding students' adoption of the ChatGPT chatbot in higher education: The role of anthropomorphism, trust, design novelty and institutional policy. *Behaviour & Information Technology*, 44(2), 315-336. <https://doi.org/10.1080/0144929X.2024.2317364>
- Qiu, L., & Benbasat, I. (2009). Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems. *Journal of Management Information Systems*, 25(4), 145-182. <https://doi.org/10.2753/MIS0742-1222250405>
- Shahzad, M. F., Xu, S., & Javed, I. (2024). ChatGPT awareness, acceptance, and adoption in higher education: The role of trust as a cornerstone. *International Journal of Educational Technology in Higher Education*, 21(1), Article 46. <https://doi.org/10.1186/s41239-024-00478-x>
- Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 3784-3803). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.findings-emnlp.320>
- Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, 12(6), 1123-1128. <https://doi.org/10.1177/1745691617708630>
- Soria, K. M. (2023). Disparities in college students' access to academic advising during the COVID-19 pandemic. *NACADA Journal*, 43(1), 17-30. <https://doi.org/10.12930/NACADA-22-27>
- Soria, K. M., Kokenge, E., Heath, C. A., Standley, E. C., Wilson, S. J. F., Connley, J. R., & Agramon, A. I. (2023). Factors associated with academic advisors' burnout. *NACADA Journal*, 43(2), 105-120. <https://doi.org/10.12930/NACADA-23-14>

- Tajvar, M., Ahmadizadeh, E., Sajadi, H. S., & Shaqura, I. I. (2024). Challenges facing international students at Iranian universities: A cross-sectional survey. *BMC Medical Education*, 24(1), Article 210. <https://doi.org/10.1186/s12909-024-05167-x>
- Tamascelli, M., Bunch, O., Fowler, B., Taeb, M., & Cohen, A. (2025). Academic advising chatbot powered with AI agent. In *Proceedings of the 2025 ACM Southeast Conference (ACMSE 2025)* (pp. 195–202). ACM. <https://doi.org/10.1145/3696673.3723065>
- Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., Bouter, L. M., & de Vet, H. C. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, 60(1), 34–42. <https://doi.org/10.1016/j.jclinepi.2006.03.012>
- Tielman, M. L., Neerincx, M. A., & Brinkman, W.-P. (2019). Design and evaluation of personalized motivational messages by a virtual agent that assists in post-traumatic stress disorder therapy. *Journal of Medical Internet Research*, 21(3), Article e9240. <https://doi.org/10.2196/jmir.9240>
- Verhagen, R. S., Neerincx, M. A., Parlar, C., Vogel, M., & Tielman, M. L. (2023). Personalized agent explanations for human-agent teamwork: Adapting explanations to user trust, workload, and performance (extended abstract). In A. Ricci, W. Yeoh, N. Agmon, & B. An (Eds.), *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023)* (pp. 2316–2318). IFAAMAS.
- Wang, S., Fatima, N., Shahbaz, M., & Asif, M. (2026). Building user trust in AI chatbots for customer service through human-like cues and perceived reliability. *Scientific Reports*, 16(1), Article 7860. <https://doi.org/10.1038/s41598-026-38179-2>
- Weng, Z., & Fu, Y. (2025). Generative AI in language education: Bridging divide and fostering inclusivity. *International Journal of Technology in Education*, 8(2), 395–420. <https://doi.org/10.46328/ijte.1056>
- Wilson, M. L., Mackay, W. E., Chi, E. H., Bernstein, M. S., Russell, D. M., & Thimbleby, H. W. (2011). RepliCHI — CHI should be replicating and validating results more: Discuss. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems* (pp. 463–466). ACM. <https://doi.org/10.1145/1979742.1979491>
- Wilson, M. L., Mackay, W. E., Chi, E. H., Bernstein, M. S., & Nichols, J. (2012). RepliCHI SIG: From a panel to a new submission venue for replication. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems* (pp. 1185–1188). ACM. <https://doi.org/10.1145/2212776.2212419>
- Wilson, M. L., Resnick, P., Coyle, D., & Chi, E. H. (2013). RepliCHI: The workshop. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (pp. 3159–3162). ACM. <https://doi.org/10.1145/2468356.2479636>
- Wilson, M. L., Chi, E. H., Reeves, S., & Coyle, D. (2014). RepliCHI: The workshop II. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems* (pp. 33–36). ACM. <https://doi.org/10.1145/2559206.2559233>
- Xiong, G., Jin, Q., Lu, Z., & Zhang, A. (2024). Benchmarking retrieval-augmented generation for medicine. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 6233–6251). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.findings-acl.372>

Zhu, X., & Yang, J. H. (2026). Exploring the role of generative AI in international students' cross-cultural adaptation: A stimulus-organism-response framework approach. *International Journal of Educational Development*, 122, Article 103568. <https://doi.org/10.1016/j.ijedudev.2026.103568>

[자기인용 · 심사본에서는 저자·제목·서지 모두 익명 처리하고 편집자에게만 제출한다] 저자 (2026). 앞선 국내 배포를 다룬 논문.

그림

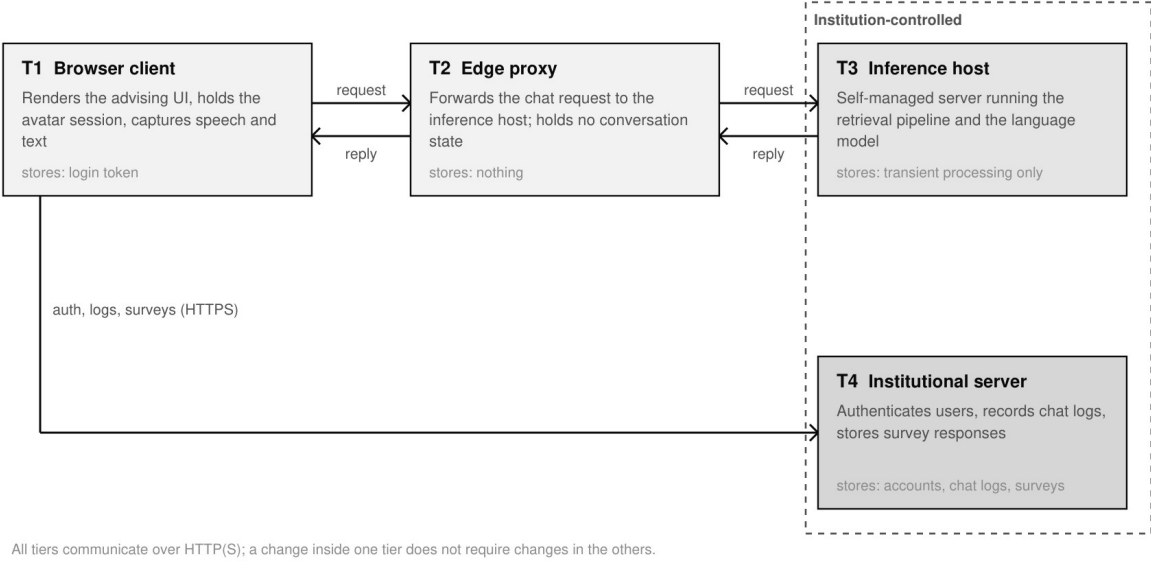


그림 1. 4 계층 시스템 아키텍처

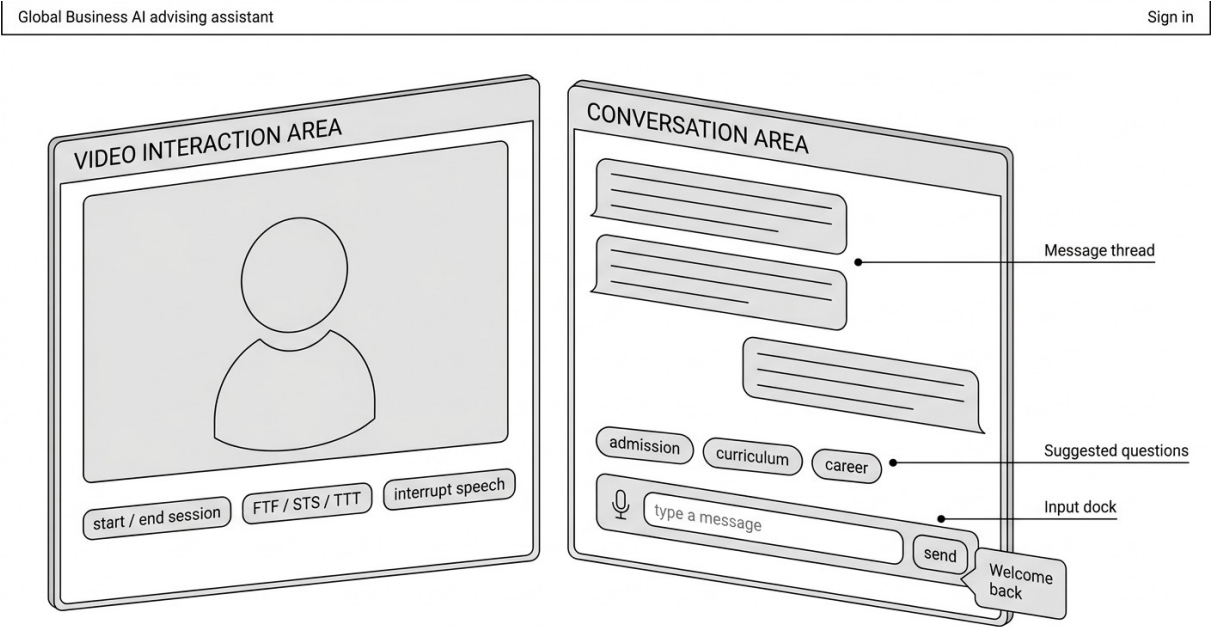
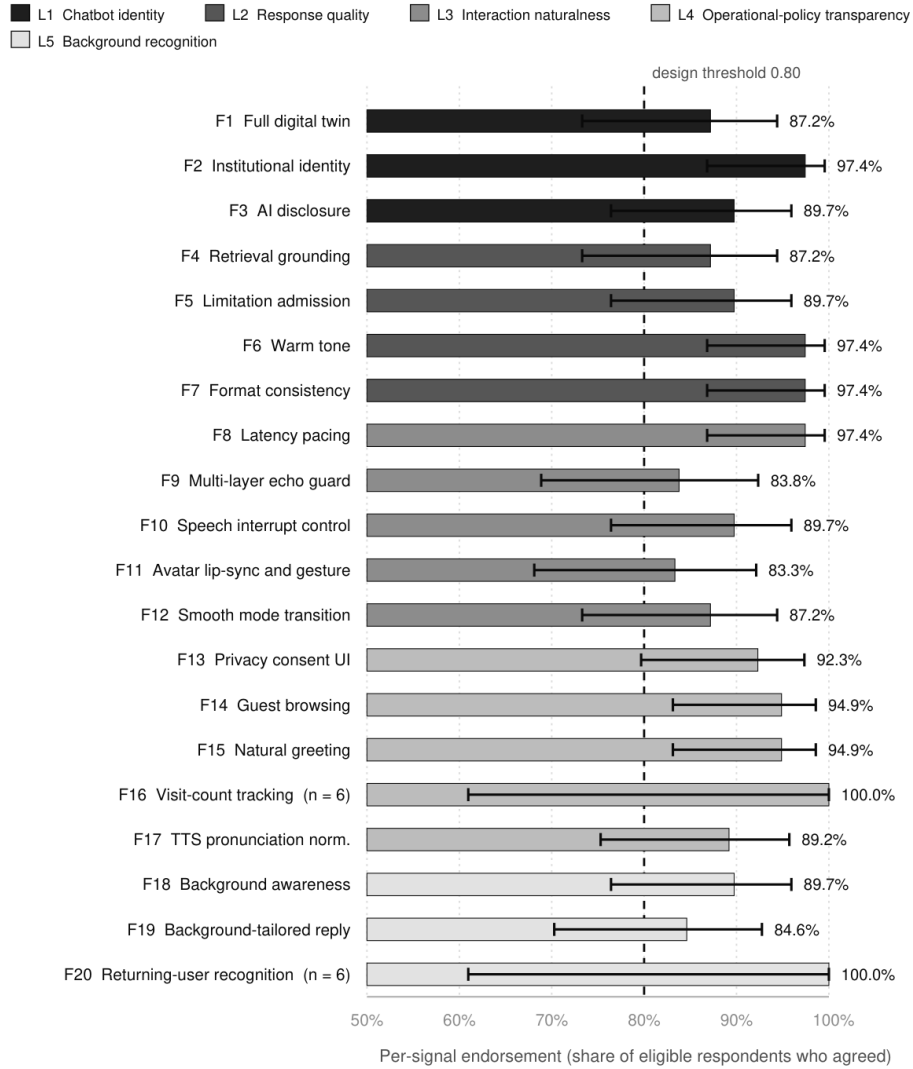
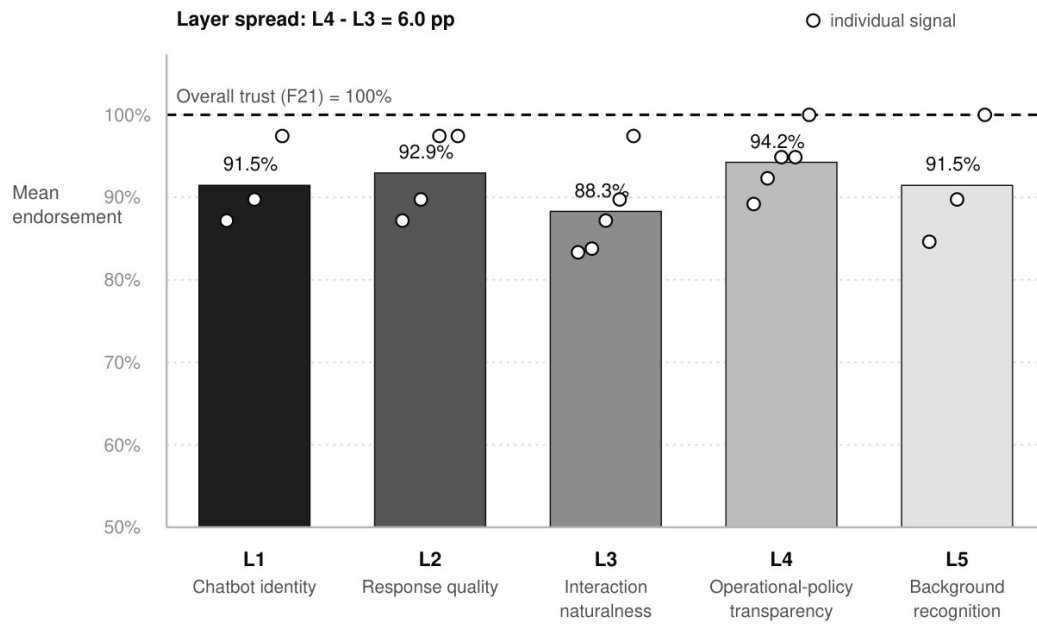


그림 2. 상담 인터페이스 구조



Bars show point estimates; horizontal lines show Wilson 95% confidence intervals. Signals marked (n = 6) are conditional; all others have n = 36-39. Only a signal whose interval lies entirely above 0.80 clears the threshold.

그림 3. 신호별 동의율과 Wilson 95% 신뢰구간



Bars show the mean endorsement of the signals in each layer; open circles show the individual signals.
 Two conditional signals (F16, F20; eligible n = 6) are included and both were endorsed by 100%, which lifts L4 and L5.

그림 4. 계층별 평균 동의율

저자 기여 (CRediT)

저자 1: 연구 착안, 방법론, 소프트웨어, 형식적 분석, 조사, 데이터 관리, 초고 작성, 시각화. 저자 2: 소프트웨어, 조사, 데이터 관리, 검토 및 편집. 저자 3: 연구 착안, 방법론, 검토 및 편집, 지도. 저자 4: 연구 착안, 방법론, 검토 및 편집, 지도, 과제 관리.

윤리 진술

본 연구는 학과 교육 도구에 대한 개선 의견을 수집하였다. 설문 진입 화면에는 약 2 분이 걸리는 익명 연구 설문이며 개인을 식별할 수 있는 자료를 수집하지 않는다는 고지와, 응답이 저자들의 연구에 쓰인다는 안내를 제시하였다. 참여는 자발적이었고 보상은 제공되지 않았으며, 응답자는 언제든지 제출하지 않고 설문을 닫을 수 있었다. 설문지는 이름·연락처를 비롯한 직접 식별자를 묻지 않았다. 분석한 39 명 가운데 6 명은 응답 시점에 서비스 계정을 보유하고 계정 참조가 함께 기록되었으나, 분석은 응답 표만으로 수행하였고 두 표를 결합하지 않았으며 본 논문과 부속 자료 어디에서도 개인이 식별되지 않는다. 성격 유형 항목은 공유본에서 제외한다. 기관생명윤리위원회 심의는 신청하지 않았다. 생명윤리 및 안전에 관한 법률 시행규칙 제 13 조에 따라 개인 식별정보를 수집·기록하지 않고 민감정보를 기록하지 않는 연구는 심의 대상이 아니며, 본 연구는 그 근거로 진행하였다.

이해관계 진술

저자들은 본 논문에 보고된 연구에 영향을 미칠 수 있는 경합하는 재정적 이해관계나 개인적 관계가 없음을 선언한다.

데이터 이용

데이터는 요청 시 제공한다. 성격 유형 항목은 공유본에서 제외한다. 39 명 가운데 8 명이 7 종으로 기재하여 그 항목만으로 대부분이 특정되기 때문이다.

연구비

본 연구는 공공·상업·비영리 부문의 어떤 기관으로부터도 특정 연구비를 지원받지 않았다.

생성형 AI 및 AI 보조 기술 사용 선언

본 원고를 준비하는 과정에서 저자들은 그림 2 의 개념도 생성에 Google Gemini 를, 원고 일부의 초안 작성과 수정, 보고 수치와 표의 정합성 확인, 참고문헌 형식 정리에 Anthropic Claude 를 사용하였다. 도구 사용 후 저자들이 내용을 검토하고 필요한 부분을 수정하였으며, 게재되는 원고의 내용에 대한 책임은 저자들에게 있다.

제목면 (별도 파일로 제출)

이중맹검이므로 본문과 분리하여 제출한다.

Trust-by-Design in an Avatar Advising Chatbot: Per-Signal Endorsement

How twenty embedded trust signals were endorsed by their users

저자 **1**^{1,†} 저자 **2**^{1,†} 저자 **3**² 저자 **4**^{2,*}

¹ 석사과정

² 교수

† 두 저자는 본 연구에 동등하게 기여하였으며 공동제일저자이다.

* 교신저자. 연락처는 별도 표제지에 있다.

[확인 필요] 저널 투고 지침은 단일 교신저자를 요구한다. 공동교신 지정 가능 여부를 편집부에 확인해야 한다.