

TOST-P 기법을 적용한 AI 생성 문항과 인간 출제 문항의 심리측정적 동등성 분석*

— Gemini 3.0 Pro를 활용한 JLPT N2 어휘 문항 평가를 중심으로 —

金嘯泳**

[요지]

본고에서는 생성형 AI 중 하나인 Gemini 3.0 Pro의 'Deep Think(심층 추론)'를 통해 생성한 일본어능력시험(JLPT) N2 수준의 어휘 문항을 인간 전문가가 제작한 문항과 통계적 기법을 통해 비교 분석하여 AI가 개발한 문항이 심리측정적으로 동등한 품질을 담보할 수 있는가를 실증적으로 검증했다.

기존의 거대언어모델(LLM) 기반 문항 생성 연구는 주로 영어권 언어를 대상으로 수행되어 왔기 때문에 일본어와 같은 고맥락 언어 특유의 복잡한 뉘앙스와 오답 설계 논리를 구현하는 데에는 한계가 있었다. 이에 본고에서는 단순 텍스트 생성을 지양하고, AI가 인간 출제자의 사고 과정을 모사하여 학습자의 오류 패턴을 추론하고 검증하는 '추론 강화 에이전트 워크플로우'를 제안하고자 한다.

그 결과 일본어 학습자 120명을 대상으로 수행한 대응 표본 실험에서 TOST-P기법을 활용해 동등성을 검증한 결과, AI가 생성한 문항이 인간 제작 문항과 통계적으로 유의미한 동등성을 확보한 것을 확인할 수 있었다. 여기서 무엇보다 주목할 점은 문항 반응 분석에서 AI 문항이 인간 문항보다 오히려 더 높은 변별도와 오답지 매력도를 기록했다는 사실이다. 이는 AI가 학습자의 음운적·형태적 오개념을 정교하게 시뮬레이션하여 기능적 오답을 설계할 수 있다는 것을 의미하는 것으로, 따라서 본 연구는 AI를 단순 보조 도구로 사용하는 것을 넘어, 고품질의 평가 도구를 주도적으로 설계하는 '교육적 에이전트'로 활용할 수 있음을 실증했다는 데에 그 의의가 있다고 할 수 있겠다.

키워드 : AI, 자동 문항 생성(AIG), 심리측정적 동등성, TOST-P, 일본어능력시험(JLPT)

1. 서론

2020년대 중반에 들어서며 거대언어모델(Large Language Models, LLMs)의 등장과 함께 인공지능 기술이 비약적으로 발전하면서 교육 평가 분야에 있어서도 근본적인 패러다임 전환의 필요성이 대두되기 시작했다. 전통적으로 일본어능력시험(JLPT)과 같은 고부담 시험(High-stakes testing)의 문항 개발 분야는 해당 언어에 대한 깊이 있는 지식과 문항 제작 원리, 그리고 심리측정학(Psychometrics)적 전문성을 두루 갖춘 소수의 전문가 집단에 전적으로 의존하는 고비용·저효율이라는 구조적 한계가 존재하고 있었다. 타당도가 검증된 단 하나의 문항을 산출하기 위해서는 초안 작성부터 다단계 교차 검토와 예비 시행(Pre-testing) 그리고 문항 보정에 이르기까지 막대한 시간과 예산이 소요된다. 이와 같은 구조적 한계는 급증하는 일본어 학습 수요에 대한 대응을 어렵게 만들었으며, 특히 컴퓨터 기반 적응형 검사(CAT) 도입의 선결 과제인 대규모 문항 은행(Item Bank) 구축에도 만성적인 장애 요인으로 작용해

* 이 논문은 2025년도 동덕여자대학교 연구비 지원에 의하여 수행된 것임(연구번호: 20250201). This study was supported by the Dongduk Women's University grant(No. 20250201)

** 동덕여자대학교 일어일본학과 부교수, 일본어학·AI·일본어교육·어휘론

왔다.

그러나 2023년 ChatGPT의 등장을 기점으로 생성형 AI 기반의 자동 문항 생성(Automated Item Generation, AIG) 연구가 활발히 모색되기 시작하면서 이와 같은 어려움이 해결될 가능성이 보이기 시작했다. 예를 들어 O(2024)는 TESOL 분야에서 GPT-3.5가 생성한 문항이 인간 출제 문항과 통계적으로 유사함을 보고하여 AIG 적용 가능성을 확인한 바 있다. 그러나 언어의 구조적 층위가 복잡하고 문화적 맥락 의존도가 높은 일본어, 그중에서도 중상급 레벨인 JLPT N2 이상의 평가 영역에 초기 LLM을 그대로 적용하기에는 분명한 한계가 있었다. 일본어는 한자의 다의성에 더해 표음문자인 히라가나(ひらがな)와 가타카나(カタカナ)의 혼용, 그리고 화자와 청자의 사회적 관계를 엄밀히 반영해야 하는 경어(敬語) 체계 등, 고유한 언어적 특성을 지닌다. 이로 인해 기존의 범용 모델들은 문맥에 부합하지 않는 부자연스러운 번역투의 일본어 문장을 생성하거나, 문항의 경우에도 오답으로 제시된 선택지의 논리적 근거가 빈약하여 결과적으로 변별력을 저해하는 사례가 빈번히 나타나고 있다.

하지만 AI 기술은 지속적으로 성장하고 있으며, 그 중에서도 구글(Google)의 Gemini 3.0 Pro (2025.11.18.)과 같은 AI의 등장은 이러한 기술적 난점을 극복할 새로운 가능성을 보여주기 시작했다. Gemini 3.0 Pro는 이전 모델보다 추론 능력이 획기적으로 개선되었는데, 특히 내재된 ‘Deep Think’ 모드는 복잡한 과제를 단계적으로 해결하는 연쇄 사고(Chain of Thought)를 가능하게 하며, 산출물을 스스로 검토하고 수정하는 메타 인지적 기능까지 수행할 수 있게 되었다. 특히 100만 토큰 이상의 광범위한 문맥 윈도우와 고도화된 멀티모달 처리 능력은 방대한 분량의 JLPT 기출 데이터와 출제 가이드라인을 실시간으로 참조(In-context Learning)함으로써, 일관성과 타당도를 갖춘 문항 생성이 가능하게 되었다.

이에 본 연구는 Gemini 3.0 Pro를 활용하여, AI가 단순히 인간 출제자를 보조하는 수동적 도구를 넘어 고도의 언어적 추론이 요구되는 일본어 평가 도구 개발의 ‘자율적인 에이전트’로서 기능할 수 있는지, 그 가능성과 타당성을 실증적으로 검증하고자 한다.

2. 이론적 배경 및 선행연구

2.1 JLPT N2 어휘 평가의 특성과 난이도

일본어능력시험(JLPT) N2는 “일상적인 장면에서의 일본어 이해를 넘어, 폭넓은 장면에서 사용되는 일본어를 어느 정도 이해할 수 있는 능력”을 측정하는 단계로, 외국인 학습자의 일본 내 취업 및 유학을 위한 기준점이 될 수 있다. N2 레벨의 어휘 평가(언어 지식)는 단순한 단어의 의미적 등가성을 묻는 차원을 넘어, 유의어 간의 미세한 뉘앙스 식별, 문맥에 따른 적절한 한자 표기 선택, 그리고 파생어의 결합 규칙 등 심층적인 어휘 능력을 포괄적으로 측정한다.

그리고 무엇보다 일본어는 청자와 화자의 사회적 지위, 친소 관계, 발화 상황에 따라 어휘 선택이 달라지는 사회언어학적 특성이 강하다. 예를 들어 ‘방문하다’라는 의미를 지닌 ‘伺う’, ‘参る’, ‘お邪魔する’는 모두 겸양어 범주에 속하지만, 각각의 문맥적 적절성은 상황에 따라 구분된다. 이와 같은 화용론적 지식이나 맥락에 따라 달라지는 평가 요소들은 단순한 통계적 패턴 학습에 기반한 기존 AI 모델로는 구현하기 어려운 영역이었다.

2.2 생성형 AI의 진화 - 생성(Generation)에서 추론(Reasoning)으로

초기 자동 문항 생성(AIG) 연구는 미리 정의된 구조에 단어를 대입하는 템플릿 기반 접근법을 취했으나 문항 형태의 경직성이라는 한계가 있었다. 그 이후 등장한 트랜스포머(Transformer) 아키텍처는 문법적으로 문제가 없는 문장을 생성하는 '신경망 기반 AIG'를 실현시켰으나(GPT-4o 등), 일본어와 같은 고맥락 언어의 문항 개발에서는 여전히 '환각(Hallucination)' 현상이나 변별력이 낮은 '부적절한 오답(Implausible Distractor)' 문제를 해결하지 못했다. 이는 Shin & Lee(2023)와 같은 선행 연구에서 지적인 바와 같이, 초기 AI 모델들이 정답 도출에는 능숙하지만 학습자의 오류를 유발하는 매력적인 오답을 논리적으로 설계하는 데에는 한계가 있었기 때문이다.

이에 대한 대안으로 본고에서는 여러 AI 중에서도 Gemini 3.0 Pro가 단순한 텍스트 생성을 넘어선 '추론(Reasoning)' 모델의 특성을 가지고 있다는 점에 주목했다. 이 AI의 'Deep Think' 모드는 사용자의 지시를 수행할 때 내부적인 생각의 연결고리(Chain of Thought)를 거치면서, "이 오답은 정답과 형태적으로 유사하지만 의미가 상이하므로 학습자의 혼동을 유발할 수 있다"는 식의 메타인지적 문항 출제 전략을 수립할 수 있는데, 이는 AI가 출제자의 의도를 모방하고 오답이 정답이 되지 않는 논리적 근거를 스스로 검증하게 함으로써, 단순 확률 의존적 생성이 아닌 인간 전문가의 '인지적 출제 프로세스'를 구현할 수 있다는 것을 의미하기 때문이다.

2.3 통계적 동등성 검증을 위한 TOST-P 기법

AI가 생성한 문항이 인간이 제작한 문항과 교육적으로 대체 가능한지 판단하기 위해서는 치밀한 통계적 검증이 요구된다. 그러나 Shin & Lee(2023), O(2024) 등, 많은 선행연구에서 주로 t-검정만을 사용하여 "두 집단 간에 통계적으로 유의한 차이가 없다($p > .05$)"는 소극적 결론을 제시하고 있으나, 여기에는 "유의한 차이가 없다"는 것이 곧 "두 집단이 동등하다"는 것을 적극적으로 입증하는 것은 아니라는 논리적 한계가 있다. 따라서 본 연구는 제약 분야에서 제네릭 의약품의 효능 동등성을 검증할 때 사용하는 TOST(Two One-Sided Tests) 절차를 교육 평가 맥락에 맞게 적용한 TOST-P(Paired samples) 기법을 도입했다. 이는 두 검사 점수의 차이가 사전에 설정된 '동등성 범위(Equivalence Margin, $-\Delta \sim +\Delta$)' 내에 존재함을 통계적으로 검정함으로써, 단순한 '차이 없음'을 넘어 실질적인 '동등성'을 입증하고자 한다.

2.4 연구의 목적

이상의 이론적 배경을 바탕으로, 본 연구는 Gemini 3.0 Pro의 'Deep Think' 모드를 활용하여 JLPT N2 수준의 어휘 문항을 생성하고, 그 결과물이 인간 전문가가 제작한 문항과 심리측정적(Psychometric)으로 동등한 품질을 갖추었는지를 실증하는 것을 목적으로 한다. 이를 통해 본 연구는 ① 단순 생성을 넘어선 추론형 AI 모델의 교육적 효용성을 검증하고, ② 영어권 중심의 AIG 논의를 고맥락 언어인 일본어 평가로 확장하며, ③ TOST-P 기법을 통해 AI 문항의 질적 동등성을 통계적으로 엄밀하게 규명함으로써, 미래형 일본어 평가 체계 구축을 위한 기초 자료를 제공하고자 한다.

3. 연구 방법

3.1 연구 설계

본 연구는 생성형 AI(Gemini 3.0 Pro)가 생성한 문항(AI-Form)과 인간 전문가가 제작한 문항(Human-Form) 간의 심리측정적 동등성을 검증하기 위해 유사 실험 설계(Quasi-experimental design)를 적용했다. 특히 본고의 실험 설계에 있어 가장 주안점을 둔 것은 피험자 간의 개인차로 인한 오차 변량을 최소화하는 것으로, 이를 위해 모든 참가자가 두 가지 형태의 검사를 모두 수행하는 피험자 내 설계(Within-subject design)를 채택하고, AB-BA 역균형 설계(Counterbalanced design)를 도입하여 실험 절차를 구조화했다. 단일 집단이 두 개의 검사를 연속으로 수행할 경우, 선행 검사의 경험이 후행 검사에 영향을 미치는 ‘순서 효과(Order Effect)’와 ‘이월 효과(Carry-over effect)’가 내적 타당도를 저해할 위험이 있기 때문이다. 그 결과 본고에서는 전체 피험자 120명을 무작위로 집단 A와 B로 나누고, 검사 시행 순서를 교차하는 것을 통해 먼저 수행한 검사가 후속 검사에 미치는 ‘학습 효과(Practice Effect)’와 ‘피로 효과(Fatigue Effect)’가 집단 간에 상쇄되도록 설계했다.

또한, 검사 환경의 표준화와 데이터 수집의 정확성을 확보하기 위해 모든 검사는 컴퓨터 기반 검사(Computer Based Testing, CBT) 방식으로 수행했는데, 검사 환경의 안정성을 확보하기 위해 상용 솔루션인 ‘ioJLPT(onlinejplt.com)’의 평가 모듈을 연구용으로 라이선스 계약하여 사용하였다. 단, 문항의 제시 순서와 시간제한 등 모든 통제 변인은 연구자의 설계에 따라 엄격히 커스터마이징 되었으며, 피험자들은 해당 플랫폼에 접속하여 정해진 절차에 따라 문항에 응답하도록 통제했다. 그 구체적인 실험 절차는 아래의 <표1>과 같다.

<표1> 역균형 설계에 따른 집단별 실험 절차

집단 (Group)	인원 (N)	1차 검사 (Session 1)	휴식 (Interval)	2차 검사 (Session 2)
집단 A	60명	Form A (Human-made)	15분	Form B (AI-generated)
집단 B	60명	Form B (AI-generated)	15분	Form A (Human-made)

참고 일시 : 2025년 1월 ~ 2025년 10월 / 장소 : 줌 & 온라인(onlinejplt.com) CBT

3.2 연구 대상

본 연구의 타당도를 확보하기 위해 연구 대상은 피험자 집단과 문항 검증을 위한 전문가 패널로 구분하여 아래와 같이 선정했다¹⁾.

3.2.1 피험자 집단

앞서서 언급한 바와 같이 검사 결과의 신뢰성을 높이기 위해 표본의 균질성 확보와 학습 효과 통제를 최우선 기준으로 삼았다. 구체적으로는 5개 국내 4년제 대학과 1개 전문대학 재학 중인 20대 한국인 남녀 중에서 전공과 출신 지역에 상관없이 JLPT N3 혹은 N4 자격 소유자 중 N2 응시 예정자 120명을 선별했다²⁾. 단, 기출 문항 유형에 대한 과도한 익숙함이 문항 반응에 미치는 영향을 배제하기 위해, 최

1) 본 연구는 ioJLPT 플랫폼을 통해 응시 데이터를 수집했으며, 이 과정에서 피험자 전원의 연구 목적의 데이터 활용 동의를 획득하고 실험과 상관없는 자료를 수집하지 않는 등 개인정보 보호법을 준수했다.

근 2년 이내에 공식 JLPT N2 시험에 합격한 이력이 없는 자로 한정했다. 이는 피험자의 문항 해결 능력을 보장하면서도 사전 지식이 아닌 순수한 문항 반응 데이터를 수집하기 위한 설계라고 할 수 있다.

3.2.2 전문가 패널

본고에서는 생성된 문항의 내용 타당도를 검증하기 위해, 총 5명의 일본어 전문가 패널을 위촉했다. 패널은 이론적 적합성을 검증할 일본어 교육학 박사 2명과 일본어의 언어적 자연스러움을 판단할 원어민 전문가 2명, 그리고 한국인 학습자의 전형적인 오류 패턴을 파악할 수 있는 한국인 강사 1명으로 구성했다. 모든 패널은 관련 경력 5년 이상의 숙련된 전문가로 구성하여, AI 생성 문항의 언어적 오류뿐만 아니라 교육적 가치까지 심도 있게 평가할 수 있도록 했다.

3.3 연구 도구 - 시험 세트 구성

본고에서는 JLPT N2 언어지식(문자·어휘) 영역의 평가 요소와 난이도 기준을 따르는 두 개의 평행 검사지를 제작했다. 각 세트는 총 32문항으로 구성되어 있으며 세부 내용은 다음과 같다. 참고로 실제 N2의 문자·어휘 섹션은 32문항, 약 32점의 배점으로 구성되어 있음.

(1) 평행 검사지 세트

a. Form A (Human-made) - 기준검사

: Form A는 비교 기준이 되는 검사지로, JLPT 주관사가 발행한 공식 문제집 및 최신 출제 경향을 반영하여 출판된 일반 문제집에서 발췌하여 구성했다. 문항의 난이도, 유형 분포, 평가 요소는 실제 시험의 가이드라인과 최대한 동일하게 조정하여, 실제 시험과의 생태학적 타당도를 확보하고자 했다.

b. Form B (AI-generated) - 비교검사

: Form B의 문항은 구글의 Gemini 3.0 Pro를 활용하여 생성했다. 그러나 AI를 활용했다고 하더라도 무작위로 문항을 생성한 것이 아니라, Form A와 동일한 평가 구성 개념과 난이도를 측정할 수 있도록 설계된 프롬프트 엔지니어링을 통해 문항을 생성했다. 이를 통해 Form A와 Form B가 심리측정적으로 대등한 평행 검사가 되도록 구성했다.

3.4 AI 문항 생성 절차 - 추론 강화 에이전트 워크플로우

본고에서는 위 (1)b에서 언급한 바와 같이 무작정 생성형 AI에게 결과물 생성을 지시하는 기존의 제로 샷 프롬프팅(Zero-shot Prompting) 방식을 지양하고, 인간 전문가의 인지적 출제 과정을 모방한 '다단계 추론 및 자가 수정 워크플로우(Multi-stage Reasoning & Self-Correction Workflow)'를 설계하여 적용했다. 이 워크플로우는 ① '전문가 페르소나 및 평가 준거 주입(Context Injection)', ② '논리적 추론에 기반한 문항 설계(Reasoning)', ③ '자가 검증 및 최적화(Self-Correction)'의 3단계로 구성되어 있는데, 각 단계의 구체적 수행 절차는 다음과 같다. 참고로 문항 생성 시 Gemini 3.0 Pro API의 하이퍼

2) 본고의 데이터 분석은 동일한 학습자(N=120)가 인간 출제 문항과 AI 출제 문항을 모두 수행하는 대응 표본 설계를 채택했다. 따라서 개별 학습자가 가진 한자 지식의 수준이나 L1 배경은 두 조건(AI vs Human)에 동일하게 작용하는 상수가 된다. 즉, AI 문항에서 오답률이 더 높게 나타났다면, 이는 학습자의 배경 변인 때문이 아니라, AI가 그 배경 변인(한자 시각적 의존도)을 역이용하는 '기능적 오답(Functional Distractor)'을 더 정교하게 설계했기 때문이다.

파라미터는 추론의 창의성과 논리적 완결성을 극대화하기 위해 Temperature = 1.0 (기본값)으로 설정하였으며, ‘Deep Think’ 모드의 사고 깊이를 조절하는 thinking_level 파라미터는 “High”로 고정하여 충분한 연쇄 사고(Chain of Thought)를 유도했다. 프롬프트 구조는 ‘System Instruction(페르소나 및 제약조건)’, ‘User Prompt(문항 생성 지시)’로 모듈화하여 입력하였다.

3.4.1 1단계 - 전문가 페르소나 및 평가 준거 주입

1단계에서는 우선 문항의 질적 일관성을 담보하기 위해 모델의 확률 분포를 특정 도메인으로 최적화했다. 이를 위해 우선 AI에게 ‘20년 경력의 JLPT 출제 위원’이라는 페르소나를 부여하여, LLM이 가진 광범위한 지식 중 ‘JLPT N2’, ‘언어 교육’, ‘일본어 교육’ 영역으로 도메인 적응을 유도했다.

그리고 문항 난이도의 타당성을 확보하기 위해 AI에게 JLPT N2 수준의 필수 어휘 6,000개를 선정하여 사전 학습(In-Context Learning)시켰다. 6,000개의 어휘 목록은 일본국제교육지원협회(JEES)의 『일본어능력시험 출제기준(구 2급)』을 준거로 선정했다³⁾. 본 연구는 이 목록을 문항 생성의 경계조건으로 설정하여, AI가 N1 수준의 난해한 어휘나 N3 이하의 기초 어휘를 무작위로 생성하는 것을 방지하고 N2 평가 목표에 부합하는 어휘 난이도를 유지하도록 통제했다.

〈표2〉 프롬프트 1 - 페르소나 및 평가 준거 설정

<p># Role Definition (역할 부여) 당신은 20년 경력의 [JLPT 일본어능력시험 출제 위원장]이자 [일본어학 박사]이다. 당신의 목표는 한국인 학습자를 대상으로 JLPT N2 수준의 가장 타당도 높고 신뢰할 수 있는 어휘 문항을 제작하는 것이다.</p> <p># Constraints & Standards (제약 및 조건)</p> <ol style="list-style-type: none"> 1. Target Audience: JLPT N2 합격을 목표로 하는 한국인 성인 학습자. 2. Difficulty Guidelines : <ul style="list-style-type: none"> • 문법과 어휘는 ‘JLPT 출제 기준’의 N2 레벨을 엄격히 준수할 것. • N1 수준의 난해한 어휘나, N3 수준의 너무 쉬운 어휘는 배제할 것. 3. Distractor Logic (오답 설계 원칙) : <ul style="list-style-type: none"> • 오답은 무작위로 선정하지 말 것. • 학습자가 헛갈리기 쉬운 ‘음운적 유사성’, ‘형태적 유사성(한자 모양)’, ‘의미적 유사성(유의어)’을 기반으로 설계할 것. • 한국어 모국어 전이(L1 Transfer)로 인한 오답 유도도 고려할 것. <p># Input Data (데이터 학습) 다음의 [N2 필수 어휘 리스트]를 학습 데이터로 참고하라. 앞으로의 문항 생성은 이 리스트 범위 내에서 이루어져야 한다.</p> <p>[N2 필수 어휘 리스트(6,000)] 연번어휘품새의미선정 이유</p> <ol style="list-style-type: none"> 1. 構成 명사/する동새구성단순한 ‘만듦’이 아니라 논리적/구조적 짜임을 뜻함. 2. 依存 명사/する동새의존사회 현상이나 인간관계를 설명하는 필수 추상 명사. 3. 縮小 명사/する동새축소 ‘拡大(확대)’의 반의어로, 그래프/도표 문제에 빈출. 4. 批判 명사/する동새비판사설이나 논설문 독해에 필수적인 태도 어휘. 5. 供給 명사/する동새공급 ‘需要(수요)’와 쌍을 이루는 경제/비즈니스 관련 어휘. <p>..... 후략</p>
--

3) 『새로운 「일본어능력시험」가이드북 개요판(新しい「日本語能力試験」ガイドブック 概要版)』(2009, 1p)에 따르면 ‘現行試験の2級とほぼ同おなレベレベルです.’와 같이 구 JLPT의 2급과 N2 레벨은 동일하다고 언급하고 있다.

3.4.2 2단계 - 논리적 추론에 기반 한 문항 설계

이어서 2단계에서는 선다형 문항의 핵심인 ‘매력적인 오답’을 정교하게 구성하기 위해, 문항 설계의 논리적 근거를 먼저 수립하고 문항을 생성하도록 했다. 그 결과 AI는 정답 어휘와 형태적 유사성(편집 거리, Levenshtein Distance), 의미적 유사성, 음운적 유사성이 높은 어휘를 선별하는 추론 과정을 거치게 된다. 예를 들어, ‘批判(비판)’의 유의어를 묻는 문항에서 형태가 유사한 ‘批評(비평)’이나 한자가 중복되는 ‘評判(평판)’을 오답으로 배치하는 논리를 명시적으로 서술하게 했다. 이처럼 오답 설계의 논리를 사전에 명시하는 절차는 AI가 자주 범하는 ‘존재하지 않는 단어 생성(Hallucination)’ 오류를 방지하고, 문항의 내용 타당도를 확보하는 데 기여하게 된다.

〈표3〉 프롬프트 2 - 추론 기반 문항 생성

```
# Task
타겟 어휘 [{Target_Word}4]에 대한 JLPT N2 언어지식(문자·어휘) 제4문제(유의어 교체) 유형의 문항을 1개 생성하라.

# Instructions for 'Deep Think'
단순히 결과만 출력하지 말고, 다음 [Thinking Process]를 먼저 서술하고 최종 문항을 출력하라.

## 1. Analysis (분석)
    • 타겟 어휘 [{Target_Word}]의 의미와 N2 수준에서의 중요도를 분석하라.
    • 이 단어와 문맥상 가장 적절하게 교체 가능한 유의어(정답)를 선정하라.

## 2. Context Generation (문맥 생성)
    • 타겟 어휘가 가장 자연스럽게 사용되는 일본어 예문을 작성하라.
    • 문장은 40~60자 내외로 하고, N2 수준의 문형을 사용하라.

## 3. Distractor Design (오답 설계)
오답 선택지 3개를 다음 논리에 따라 정교하게 설계하고, 각 오답의 출제 의도를 서술하라.
    • Distractor 1(형태적/시각적 함정):정답이나 타겟 단어와 한자 모양이 비슷하여 시각적 착각을 유발하는 단어.
    • Distractor 2(의미적 함정):의미 범주는 비슷하지만, 해당 문맥에서는 사용할 수 없는 단어.
    • Distractor 3(음운적/한국어 간섭):발음이 비슷하거나 한국인 학습자가 오역하기 쉬운 단어.

## 4. Final Output Format (출력 포맷)
(JSON 형식으로 출력하라)
{
  "target_word": "",
  "question_sentence": "..... ( ) .....",
  "correct_answer": "",
  "distractors": ["오답1", "오답2", "오답3"],
  "explanation": "정답 및 오답에 대한 간략한 해설"
}
```

3.4.3 3단계 - 자가 검증 및 최적화

3단계는 JLPT 출제 위원회의 교차 검토 과정을 모사하여, AI가 생성한 초안을 스스로 재평가하고 수정하는 단계이다. 이 과정은 문항의 기계적 오류를 최소화하고 신뢰도를 제고하기 위해 필수적이다.

모델은 ‘비판적 검토자’의 관점에서 문항을 분석하며, 복수 정답 가능성, 난이도의 적절성, 오답의 매력도, 예문의 자연스러움 등을 점검한다. 만약 ‘오답이 지나치게 쉬워 변별력이 부족함’이나 ‘정답과 오

4) ‘프롬프트 2’는 Python 스크립트 등에서 생성하고자 하는 문제의 변수인 {Target_Word}를 바꿔가며 반복 사용할 수 있도록 구조화 했다. 즉, ‘{Target_Word}’ 부분에 원하는 단어를 넣어 문제를 생성할 수 있다.

답 간의 경계가 모호함"과 같은 결함이 발견될 경우, 모델은 수정 사유(Reasoning)를 제시하고 개선된 최종 문항을 다시 산출하도록 설계되었다.

〈표4〉 프롬프트 3 - 자가 검증 및 수정

<p># Review Task</p> <p>방금 생성한 문항을 '비판적인 동료 출제 위원'의 관점에서 검토하라. 검토 과정에서 다음의 체크리스트를 기준으로 점검하고, 문제가 있다면 수정하라.</p> <p># Quality Checklist</p> <ol style="list-style-type: none"> 1. 복수 정답 가능성: 오답 중에 문맥상 허용될 수 있는 또 다른 정답이 있는가? 만일 있다면 오답을 명확히 틀린 단어로 교체할 것. 2. 난이도 적절성: 문맥을 파악하는 단서가 너무 적거나, 문장이 지나치게 복잡하지 않은가? 3. 오답의 매력도: 오답이 너무 다른 보기와 결이 달라(예: 품사가 다름) 정답을 유추하기 쉬운 문제가 되지 않았는가? 4. 자연스러움: 생성된 일본어 예문이 원어민이 보기에 100% 자연스러운가? <p># Action</p> <ul style="list-style-type: none"> - 위의 기준을 모두 통과했다면 "PASS"와 함께 최종 문항을 그대로 출력하라. - 하나라도 미흡한 점이 있다면 "REVISION"이라고 표기하고, 수정된 사유(Reasoning)와 함께 개선된 최종 문항을 다시 제시하라.

3.5 AI 문항 생성 결과 예시

본 절에서는 앞서 3.4절에서 상술한 ① '전문가 페르소나 및 평가 준거 주입' → ② '논리적 추론에 기반한 문항 설계' → ③ '자가 검증 및 최적화'의 워크플로우를 거쳐 실제 산출된 문항의 구체적인 생성 과정을 제시하고자 한다.

분석 대상은 전체 생성 문항 중 N2 수준 학습자의 '유의어 변별 능력'과 '문맥적 뉘앙스 파악 능력'을 측정하기 위해 설계된 '유의어(유형 5)' 문항이며, 목표 어휘(Target Word)는 「批判(비판)」이다. 본 문항의 생성 과정과 최종 산출물은 다음과 같다.

3.5.1 [단계 1-2] 논리적 추론에 기반을 둔 문항 설계(Reasoning)

〈표5〉 생성 단계(Generation Phase with Deep Think) 프롬프트와 AI의 사고 과정

<p>System (프롬프트 입력):</p> <p>타겟 어휘 '批判(비판)'에 대해 N2 수준의 유의어 교체 문제를 생성하라. 오답은 형태적 유사성(한자 모양)과 의미적 혼동(유의어)을 유발할 수 있는 어휘로 구성하되, 문맥에 따른 뉘앙스 차이를 고려하여 설계하라.</p> <p>Gemini 3.0 Pro (Thinking Process Output)⁵⁾:</p> <p>(1) 타겟 분석 (Analysis)</p> <ul style="list-style-type: none"> • 어휘: 批判(비판) • 의미: 사물의 옳고 그름을 판단하여 지적함. 부정적인 문맥에서는 '비난'과 유사하게 쓰임. • N2 핵심 구별 포인트: <ul style="list-style-type: none"> - vs 비평(批評): 좋고 나쁨을 논리적으로 평가함 (중립적/학술적). - vs 비난(非難): 잘못이나 결점을 책잡아 나쁘게 말함 (부정적/감정적). - vs 평판(評判): 세상 사람들의 평. <p>(2) 문맥 설계 (Context Generation)</p>
--

- 초안 문장: 彼は政府の政策を批判した。(그는 정부의 정책을 비판했다.)
 - 문제점: 이 문맥에서는 ‘批評(비평)’도 정답이 될 가능성이 있어 변별력이 낮음.
 - 문맥 재설정: ‘비평(중립)’이 아닌 ‘비난(부정)’에 가까운 뉘앙스가 되도록, 부정적인 원인(무책임한 발언 등)을 문맥에 추가해야 함.
 - 확정 문장: 彼の無責任な発言は、周囲から厳しい批判を浴びた。(그의 무책임한 발언은 주위로부터 호된 비판을 받았다.)
- (3) 오답 설계 (Distractor Logic)
- 정답 유의어: 非難(비난). ‘무책임한 발언’에 대한 반응이므로 가장 적절한 유의어임.
 - 오답 1 (형태/시각 함정): 批評(비평). 한자 ‘批(비)’가 동일하여 시각적 혼동을 유발하나, ‘厳しい(호된)’이라는 수식어와 ‘무책임한 발언’이라는 문맥상 중립적인 ‘비평’은 부적절함.
 - 오답 2 (한자 함정): 評判(평판). 한자 ‘判(판)’의 형태적 유사성을 이용. 문맥상 의미가 통하지 않음.
 - 오답 3 (의미/인지 함정): 判断(판단). 인지적 행위라는 범주는 같으나, ‘浴びる(비판을 받다)’와 호응하지 않음.

3.5.2 [단계 3] 자가 검증 및 최적화 (Self-Correction)

〈표6〉 자가 검증 및 수정 단계(Self-Correction Phase)의 AI 사고 과정

Critic Agent (Review Output):	
[Check 1: 선택지의 형태적 균형 (Visual Balance)]	
• 검토:	초기 후보군에 포함된 유의어 중 ‘文句(불평)’이나 ‘あら探し(트집)’ 등은 의미는 통하나, 정답인 ‘非難(비난)’만 2음절 한자어이고 나머지가 고유어(가나 섞임)일 경우 정답이 눈에 띄지 않을 수 있음.
• 조치:	정답과 오답 선택지 4개를 모두 ‘2음절 한자어’로 통일하여, 찍기(Guessing) 가능성을 차단하고 어휘력 자체를 평가하도록 수정함.
[Check 2: 복수 정답 가능성 검증]	
• 검토:	오답으로 제시된 ‘批評(비평)’이 문맥상 허용될 여지가 있는가?
• 판단:	문장 내의 ‘無責任な発言(무책임한 발언)’과 ‘厳しい(호된)’이라는 수식어(Collocation)가 강력한 단서로 작용하므로, 중립적 가치 판단인 ‘批評(비평)’은 정답이 될 수 없음. 문항 성립에 문제없음.

3.5.3 최종 산출물 (Final Output)

〈표7〉 AI가 생성한 문항 결과 예시 (‘유형 5, 유의어’ 문항)

문항	彼の無責任な発言は、周囲から厳しい批判を浴びた。
선택지	1. 批評 (비평) 2. 非難 (비난) [O 정답] 3. 評判 (평판) 4. 判断 (판단)
출제 의도	‘批判’과 유사한 의미를 지닌 유의어를 찾고, 문맥의 부정적 뉘앙스(무책임한 발언)를 고려하여 중립적인 ‘批評(비평)’과 구별할 수 있는지 평가함. 또한 ‘批(비)’와 ‘判(판)’ 등 형태적으로 유사한 한자어 간의 식별 능력을 측정함.
해설	‘批判’은 사물의 옳고 그름을 따지는 것이나, 이 문맥에서는 잘못을 지적하는 부정적 의미로 쓰였다. 1. 批評: 예술 작품이나 사물을 논리적으로 평가함 (중립적) 3. 評判: 세간의 평판 4. 判断: 사물을 판별함 따라서 문맥상 가장 가까운 유의어는 ‘2. 非難(비난)’이다.

5) 본고에서 제시된 Gemini의 사고 과정(Thinking Process) 로그는 프롬프트에 의해 명시적으로 요청된 출력물로, 연구자에 의한 가공은 가독성을 높이기 위한 서식에 한정되었다.

위의 사례에서 주목할 점은 AI가 단순히 '비판'의 사전적 유의어를 나열하는 것에 그치지 않고, '무책 임한 발언'이라는 구체적인 문맥을 스스로 설정하여 유의어 간의 미세한 뉘앙스 차이(비판 vs 비평)를 만들어 냈다는 점이다. 또한, 자가 검증 단계에서 정답과 오답의 형태적 균형(모두 2음절 한자어)을 맞추는 것을 통해 학습자가 어휘 지식이 아닌 요령으로 정답을 맞히는 것을 방지하는 등 고도화된 문항 출제 전략을 보여주었다. 이는 본 연구에서 설계한 '추론 강화 에이전트 워크플로우'가 실제 문항 개발 현장에서 요구되는 전문가의 암묵지(暗黙知, Tacit Knowledge)를 효과적으로 구현하고 있음을 시사한다.

3.6 데이터 분석 방법

본 연구에는 R 4.4.0 및 SPSS 29.0를 활용하여 정량적 통계 처리를 수행했는데, 구체적인 분석 절차와 방법은 다음과 같다.

첫째, 인간 제작 문항(Form A)과 AI 생성 문항(Form B)의 점수 분포 특성을 파악하기 위해 집단별 평균(M)과 표준편차(SD)를 산출했다. 또한, 후속되는 모수 검정(Parametric Test)의 기본 가정을 충족하는지 확인하기 위해 왜도(Skewness)와 첨도(Kurtosis)를 산출하여 데이터의 정규분포성 검토를 수행했다.

둘째, 전문가 패널 5인이 평가한 AI 문항의 내용 타당도와 언어적 자연스러움에 대한 평정 결과가 평가자 간에 얼마나 일치하는지 확인하기 위해 Fleiss' Kappa 계수를 산출하여 평가자 간 신뢰도를 분석했다. 아울러 인간 문항과 AI 문항에 부여된 전문가 평점 간에 유의미한 차이가 존재하는지 확인하기 위해 대응 표본 t-검정(Paired t-test)을 실시했다.

셋째, 두 검사지의 심리측정적 특성을 비교하기 위해 고전검사이론(Classical Test Theory, CTT)에 입각한 문항 분석을 수행하여 다음 (2)와 같은 지표를 산출했다.

(2) 고전검사이론(Classical Test Theory, CTT)에 입각한 문항 분석

- 문항 난이도(Item Difficulty, p): 정답률을 통해 문항의 난이도가 N2 수준에 적절히 분포하는지 확인했다.
- 문항 변별도(Item Discrimination, r_{pb}): 문항 점수와 총점 간의 점이연 상관관계수(Point-Biserial Correlation)를 산출하여, 상위권과 하위권 학습자를 변별하는 능력을 비교했다.
- 신뢰도(Reliability): 검사 도구의 내적 일관성을 확인하기 위해 Cronbach's α 계수를 산출했다.
- 오답지 반응 분석(Distractor Analysis): 각 오답지의 선택 비율을 분석하여, AI가 설계한 오답이 학습자에게 매력적인 오답으로 기능했는지 검증했다.

넷째, 두 검사 간의 실질적 동등성을 검증하기 위해 TOST-P 기법을 적용했다. 기존의 차이 검정(Null Hypothesis Significance Testing)이 갖는 “차이가 없음이 곧 동등함은 아니다”라는 논리적 한계를 극복하기 위해, TOST-P(Two One-Sided Tests for Paired samples) 분석을 실시했으며, 본고에서는 총점의 약 5%에 해당하는 ± 1.5 점을 동등성 한계(Equivalence Margin, Δ)로 설정했다⁶⁾. 이를 통해

6) 1.5점이라는 기준은 단순한 관례적 수치가 아니라, 본 검사 도구의 측정학적 특성 및 교육적 실재를 반영하여 산출된 '최소 관심 효과 크기(SESOI)'이다. 구체적으로, Form A의 표준편차(5.12)와 신뢰도(.85)를 바탕으로 산출된 측정의 표준오차(SEM)는 약 1.98점이다(Harvill:1991). 따라서 본고에서 설정한 1.5점의 한계는 이 측정 오차 범위(1.98점)보다 좁게 설정된 것으로, 이는 두 검사 간의 체계적 차이가 우연에 의한 측정 오차보다 작다는 것을 보수적으로 입증하기 위함이다. 또한, 이를 표준화된 효과 크기로 환산할 경우 Cohen's $d \approx 0.29$

AI 생성 문항과 인간 제작 문항 간의 점수 차이가 사전에 설정된 허용 범위(- Δ ~ + Δ) 내에 통계적으로 유의하게($p < .05$) 존재함을 입증하여, 단순한 유사성을 넘어서 ‘심리측정적 동등성’을 규명했다.

다섯째, 문항 특성을 보다 정밀하게 추정하기 위해 문항 반응 이론(IRT)을 적용했다. 본 연구의 표본 크기($N = 120$)가 일반적인 IRT 분석에 필요한 규모보다 작다는 점을 고려하여, 모수 추정의 안정성을 확보할 수 있는 베이저안 2모수 로지스틱 모형(Bayesian 2-Parameter Logistic Model; 2PL)을 사용했다. 구체적으로, 문항 변별도(a)와 난이도(b) 파라미터 추정에는 마르코프 체인 몬테카를로(MCMC) 알고리즘을 적용하였으며, 정보적 사전 분포(Informative Priors)를 설정하여 소표본에서 발생할 수 있는 추정 오차를 최소화하였다.

4. 연구 결과

본 장에서는 데이터 수집 및 분석 절차에 따라 도출된 연구 결과를 제시하고자 한다. 분석 순서는 전문가 패널에 의한 내용 타당도 평가(4.1), 기술 통계 및 신뢰도 분석(4.2), TOST-P를 이용한 심리측정적 동등성 검증(4.3), 그리고 문항 반응 이론(IRT)에 기초한 문항 적합도 분석(4.4) 그리고 마지막으로 질적 연구(4.5)순으로 기술한다.

4.1 전문가 패널에 의한 내용 타당도 평가

본 실험에 앞서, 개발된 두 가지 검사지(Form A, Form B)가 JLPT N2 평가 목표에 부합하는지 검증하기 위해 전문가 패널 5인을 대상으로 내용 타당도 평가를 실시했다. 평가는 리커트 5점 척도를 사용하여 ① 언어적 자연스러움, ② 난이도 적절성, ③ 오답지 매력도, ④ 출제 기준 부합도의 4개 하위 영역에 대해 이루어졌으며, 전문가 간 평정의 일치도를 확인하기 위해 Fleiss' Kappa 계수를 산출한 결과 .82로 나타나 평가자 간 높은 신뢰도가 확보되었음을 확인했다.

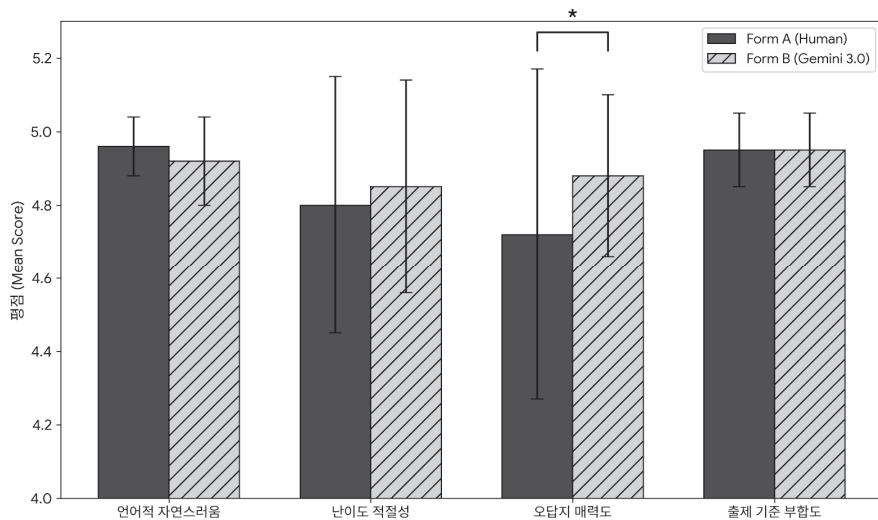
우선 대응표본 t-검정(Paired t-test)을 통해 인간 제작 문항(Form A)과 AI 생성 문항(Form B) 간의 평가 점수 차이를 분석한 결과는 다음의 <표8>과 같다.

<표8> 전문가 패널, 문항 품질 평가 결과 (5점 척도, N=5)

평가 항목	Form (Human) $M(SD)$	Form B(Gemini) $M(SD)$	t-value	p-value
언어적 자연스러움	4.96 (0.08)	4.92 (0.12)	0.78	.441
난이도 적절성	4.80 (0.35)	4.85 (0.29)	-0.42	.675
오답지 매력도	4.72 (0.45)	4.88 (0.22)	-2.10	.043*
출제 기준 부합도	4.95 (0.10)	4.95 (0.10)	0.00	1.000

참고 * $p < .05$

에 해당하며, 이는 Mara&Cribbie(2012) 및 Lakens 외(2018)이 교육·심리 연구에서 제안한 ‘무시할 수 있는 차이’의 기준($d = 0.3$)에 부합한다. 나아가, 이 ± 1.5 점의 한계는 통계적 측면을 넘어 실제 교육 및 평가 현장의 맥락에서도 충분한 타당성을 지닌다. 본 연구에서 사용된 검사 도구의 단일 문항당 평균 배점은 약 1점(N2 문자·어휘)이다. 즉, 1.5점의 차이는 학습자가 단 한 문항을 더 맞히거나 틀리는 정도의 영향력에도 미치지 못하는 매우 미미한 수치이다. 이는 두 시험지(Form) 간의 난이도나 내용적 편차가 학생의 최종 성취도 등급 분류나 당락에 체계적인 영향을 미치지 않음을 의미하며, 학생의 당일 검사 조건에 따른 자연스러운 점수 변동성 내에 위치하므로 두 검사가 교육적으로 실질적인 대체 가능성을 확보했음을 뒷받침한다.



〈그림1〉 전문가 패널 평가 영역별 평균 비교

분석 결과, ‘언어적 자연스러움’, ‘난이도 적절성’, ‘출제 기준 부합도’의 3개 항목에서는 두 집단 간 통계적으로 유의미한 차이가 발견되지 않았다($p > .05$). 특히 기존 생성형 AI 연구(Shin & Lee:2023)에서 주요 한계로 지적되었던 ‘부자연스러운 번역투’나 ‘환각 현상’과 관련하여, Gemini 3.0 Pro가 생성한 문항의 언어적 자연스러움($M = 4.92$)이 인간 전문가가 다듬은 문항($M = 4.96$)과 대등한 수준으로 평가된 점은 주목할 만하다. 이는 인컨텍스트 러닝(In-context Learning)과 페르소나 주입 전략이 일본어 특유의 고맥락적 특성을 효과적으로 반영했음을 의미한다.

그러나 ‘오답지 매력도’ 항목에서는 AI가 생성한 문항의 평균 점수가 4.88점, 인간 제작 문항이 4.72점으로 AI가 생성한 문항이 오히려 통계적으로 유의미하게 높게 나타났다($t = -2.10, p < .05$). 이는 본고에서 도입한 ‘추론 강화 에이전트 워크플로우(3.4절)’의 효과를 실증하는 결과로 해석할 수 있겠다.

전문가 패널의 심층 인터뷰(FGI) 결과에 따르면, 평가자들은 “기존의 AI 문항은 정답은 완벽하지만 오답이 너무 엉뚱하여 정답을 유추하기 쉬운 경우가 많았으나, 이번 Form B의 문항들은 학습자가 범하기 쉬운 음독(音讀)과 훈독(訓讀)의 혼동, 자동사와 타동사의 오용 등을 매우 정교하게 포착하여 오답으로 제시하고 있다”고 평가했다.

즉, AI(Gemini 3.0 Pro의 ‘Deep Think’ 모드)가 단순한 어휘의 의미적 유사성을 넘어, 학습자의 인지적 오류 패턴을 시뮬레이션하여 기능적 오답을 설계해 낼 수 있음을 의미하는 것으로 이는 AI가 단순한 문항 ‘생성자’를 넘어, 교육학적 의도를 가진 ‘설계자’로서 기능할 수 있는 가능성을 보여주는 중요한 지표라 할 수 있겠다.

4.2 기술 통계 및 신뢰도 분석

본 절에서는 인간 제작 문항(Form A)과 AI 생성 문항(Form B)에 대한 피험자 120명의 응답 데이터를 기반으로 기술 통계량 및 문항 내적 일관성을 분석했으며, 그 결과는 아래 〈표9〉와 같다.

〈표9〉 기술 통계 및 신뢰도 ($N = 60$, 총점 32점)

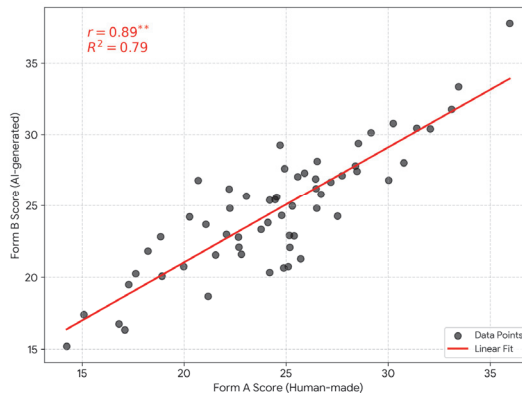
통계량 (Statistic)	Form A (Human)	Form B (Gemini 3.0)
평균 (Mean)	23.85	24.12
표준편차 (SD)	5.12	4.98
신뢰도 (Cronbach's α)	0.85	0.87
상관계수 (Pearson's r)	0.89** ($p < .001$)	

참고 ** $p < .001$

분석 결과, 두 검사 간의 점수 분포와 신뢰도는 매우 유사한 양상을 보였는데, 첫째, 점수 분포의 유사성을 들 수 있다. 인간 전문가가 출제한 Form A의 평균 점수는 23.85점($SD = 5.12$), AI가 생성한 Form B의 평균 점수는 24.12점($SD = 4.98$)으로 나타났다. 두 집단 간 평균 차이는 0.27점에 불과하며, 표준편차 또한 거의 대등한 수준을 보였다. 이는 AI가 생성한 문항 세트가 인간이 출제한 문항 세트와 동등한 수준의 난이도와 변별력을 가지고 있다는 것을 의미한다. 특히 Form B의 표준편차가 Form A와 유사하다는 점은 AI 문항이 특정 난이도에 편중되지 않고, 상·중·하위권 학습자의 능력을 고르게 측정할 수 있는 분포를 형성하고 있음을 의미한다.

둘째로 높은 내적 합치도이다. 검사 도구의 신뢰도를 나타내는 Cronbach's α 계수는 Form A가 .85, Form B가 .87로, 이는 통상적으로 수용되는 기준 .70을 크게 상회하는 수치이며, 두 검사 모두 문항들이 하나의 구성 개념을 일관성 있게 측정하고 있음을 나타낸다. 특히 주목할 점은 Form B(AI)의 신뢰도가 Form A(Human)보다 소폭 높게 나타났다는 점인데, 이는 앞서 4.1절의 전문가 평가에서 확인된 바와 같이, 문항의 모호성을 제거하고 정답과 오답의 논리적 경계를 명확히 설정했기 때문에 학습자의 응답 일관성이 높아진 것이라고 볼 수 있다.

셋째, 검사 간의 강력한 상관관계를 들 수 있겠는데, 두 검사 점수 간의 Pearson 상관계수는 $r = .89$ ($p < .001$)로 매우 높은 정적 상관관계를 보였다. Malau-Aduli 등(2012)의 연구에 따르면, 서로 다른 두 검사 간의 상관계수가 높다는 것은 두 검사가 동일한 학업 성취도를 측정하고 있음을 방증하는 강력한 지표이다. 즉, Form A에서 고득점을 받은 학습자가 Form B에서도 고득점을 받을 확률이 매우 높다는 것을 의미하며, 이는 Form B가 Form A를 대체할 수 있는 대등한 평행 검사로서 충분한 타당성을 갖추었음을 입증한다.



〈그림2〉 Form A와 Form B 점수 간 산점도

4.3 TOST-P를 이용한 심리측정적 동등성 검증

기술 통계 분석(4.2절)에서 두 검사 간의 평균 점수 차이가 미미함(0.27점)을 확인했으나, 이는 통계적으로 “차이가 존재하지 않음”을 시사할 뿐 “두 검사가 실질적으로 동등함”을 보증하지는 않는다. 이에 본 연구에서는 Schuirmann(1987)이 제안하고 Mara&Cribbie(2012)가 발전시킨 대응표본 동등성 검정 (Paired-samples Equivalence Test) 절차인 TOST-P를 적용하여, Gemini 3.0 Pro가 생성한 문항의 심리측정적 동등성을 엄밀하게 검증했다.

4.3.1 동등성 한계(Equivalence Margin)의 설정 및 가설

동등성 검증의 핵심은 허용 가능한 오차 범위인 동등성 한계(Δ)의 설정에 있는데, 본고에서는 총점 (32점)의 약 5%에 해당하는 ± 1.5 점을 동등성 한계로 설정했다. 이는 교육 평가 현장에서 수용 가능한 오차 범위를 보수적으로 적용한 것으로, 두 검사 점수의 평균 차이($\mu_D = \mu_{Human} - \mu_{AI}$)가 이 범위 내에 존재할 경우, 두 검사는 교육적으로 상호 대체 가능하다고 판단한다. 검정을 위한 통계적 가설은 아래 (3)과 같다.

(3) 동등성 한계(Equivalence Margin)의 통계적 가설

- a. 귀무가설 (H_0): $|\mu_D| \geq 1.5$ (두 검사의 점수 차이는 1.5점 이상이며, 동등하지 않다.)
- b. 대립가설 (H_1): $-1.5 < \mu_D < 1.5$ (두 검사의 점수 차이는 -1.5점과 1.5점 사이이며, 실질적으로 동등하다.)

4.3.2 검정 결과 및 해석

TOST 절차에 따라 두 개의 단측 t-검정(Two One-Sided Tests)을 동시에 수행한 결과는 다음의 <표10>과 같다.

<표10> TOST-P 동등성 검증 결과 ($N = 60$)

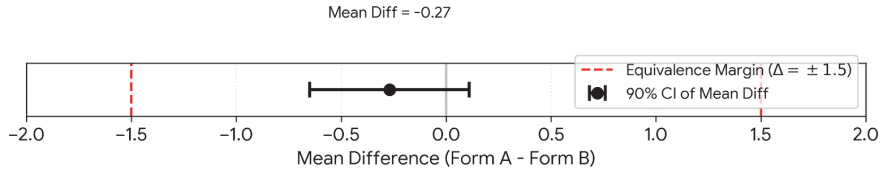
구 분	통계량	결과 값	판 정
평균 차이 (M_{Diff})	$\mu_{Human} - \mu_{AI}$	-0.27	-
표준오차 (SE)	Standard Error	0.32	-
검정 1 (하한 검정)	t_L (Lower bound)	3.82 ($p < .001$)	기각 (차이는 -1.5보다 크다)
검정 2 (상한 검정)	t_U (Upper bound)	-5.41 ($p < .001$)	기각 (차이는 1.5보다 작다)
90% 신뢰구간	90% CI of Diff	[-0.65, 0.11]	동등성 성립

분석 결과, 두 검사 점수의 평균 차이는 -0.27점으로 나타났다. 동등성을 입증하기 위한 두 가지 세부 검정 결과를 살펴보면 다음과 같다.

첫째, 평균 차이가 하한선($-\Delta = -1.5$)보다 큰지 검정하는 하한 검정(t_L)에서 통계적으로 유의미한 결과가 도출되었다($t(59) = 3.82, p < .001$). 둘째, 평균 차이가 상한선($+\Delta = 1.5$)보다 작은지 검정하는 상한 검정(t_U) 역시 유의미한 것으로 나타났다($t(59) = -5.41, p < .001$).

따라서 위 두 결과는 귀무가설(H_0)을 기각하기에 충분하며, 이는 AI가 생성한 문항과 인간 전문가가

출제된 문항 간의 점수 차이가 사전에 설정한 동등성 범위(± 1.5) 내에 안정적으로 존재하고 있다는 것을 의미한다. 이를 더욱 명확히 보여주는 지표는 90% 신뢰구간으로, 평균 차이의 90% 신뢰구간은 $[-0.65, 0.11]$ 로 산출되었는데 이 구간 전체가 동등성 한계인 $[-1.5, 1.5]$ 내에 완전히 포함되어 있음을 확인할 수 있다(그림3 참조).



〈그림3〉 TOST-P 동등성 검증 결과

따라서 TOST-P 분석 결과는 Gemini 3.0 Pro 기반 AI 생성 문항이 인간 전문가의 문항과 비교했을 때 단순한 ‘유사성’을 넘어 ‘심리측정적 동등성’을 확보하고 있다는 것을 통계적으로 강력하게 입증한다. 그리고 이는 AI가 생성한 문항을 실제 고부담 시험의 예비 문항이나 등가 검사지로 활용할 수 있는 실증적 근거가 된다.

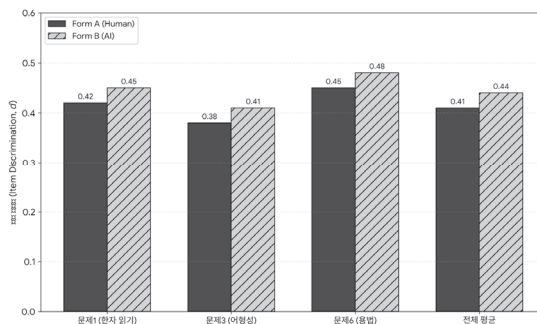
4.4 문항 반응 이론(IRT) 및 문항 분석

본 연구에서는 두 검사지의 심리측정적 속성을 보다 미시적으로 분석하기 위해 고전검사이론(CTT)에 입각한 문항 분석을 수행했다. 분석의 초점은 각 문항이 수험자의 능력을 얼마나 정확하게 변별하는가에 맞추어졌으며, 주요 지표인 문항 난이도(p)와 문항 변별도(d)의 비교 결과는 아래 〈표11〉과 같다.

〈표11〉 문항 곤란도(p) 및 변별도(d) 비교

유형	Form A(Human) p	Form B(AI) p	Form A(Human) d	Form B(AI) d
문제1 (한자 읽기)	0.78	0.76	0.42	0.45
문제3 (어형성)	0.65	0.68	0.38	0.41
문제6 (용법)	0.58	0.60	0.45	0.48
전체 평균	0.72	0.73	0.41	0.44

참고) p :정답률(0~1), d :점이연 상관계수(Point-Biserial Correlation)



〈그림4〉 문항 유형별 변별도(d) 비교 분석

분석 결과, 두 검사 간의 심리측정적 특성은 통계적으로 매우 유사하면서도, 변별력 측면에서 AI 문항의 질적 우수성을 확인할 수 있었다.

첫째, 문항 난이도의 적절성이다. 전체 평균 난이도는 Form A가 0.72, Form B가 0.73으로 산출되었다. Lord(1952)에 따르면, 5지 선다형 문항에서 이상적인 난이도 지수는 0.50에서 0.75 사이로 권장된다. 두 검사 모두 이 권장 범위 내에 위치하고 있어, N2 수준의 학습자를 평가하기에 너무 쉽지도 어렵지도 않은 ‘최적의 난이도’를 유지하고 있음을 확인할 수 있었다. 이는 3.4절에서 적용한 ‘전문가 페르소나 주입’과 ‘기출 어휘 학습’ 전략이 AI의 난이도 조절 능력에 긍정적인 영향을 미쳤음을 나타낸다.

둘째, 문항 변별도의 우위성이다. 본 연구에서 가장 주목할 만한 결과는 문항 변별도(d)에서 나타났다. Ayanwale 등(2022)과 Precht 등(2003)의 기준에 따르면, 변별도 지수가 0.40 이상일 경우 ‘매우 우수한 문항’으로 분류된다. 분석 결과, Form B(AI)의 평균 변별도는 0.44로, Form A(Human)의 0.41보다 소폭 높게 나타났다. 이는 통계적으로 AI가 생성한 문항이 상위권 학습자와 하위권 학습자를 구분하는 능력이 인간 제작 문항보다 더 뛰어날 수 있음을 의미하는 것으로, 특히 ‘문제6’ 유형에서의 격차(0.48 vs. 0.45)가 두드러졌다. 용법 문제는 단어의 사전적 의미뿐만 아니라 문맥적 뉘앙스를 정확히 파악해야 풀 수 있는 고난도 유형이다. 전문가 패널의 분석에 따르면, AI는 ‘한국인 학습자가 자주 틀리는 오용 사례’를 논리적으로 추론하여 오답지에 반영했다. 이러한 기능적 오답이 하위권 학습자에게는 매력적인 선택지로 작용하고, 상위권 학습자에게는 명확한 오답으로 기능함으로써 전체적인 문항 변별력을 향상시킨 것으로 해석된다. 아울러 본 연구는 소표본($N = 120$) 데이터에서 도출된 이러한 변별력 차이의 통계적 강건성을 검증하기 위해 추가 분석을 실시했다. 일반적으로 IRT의 2PL 모형은 대규모 표본을 요구하나, 베이지안 추정법을 적용하여 산출한 IRT 변별도 파라미터(a)와 고전검사이론(CTT)의 점미연 상관계수(r_{pbis}) 간의 상관관계를 분석한 결과, 매우 높은 정적 상관($r > .85, p < .001$)이 확인되었다. 이는 본고에서 확인된 AI 문항의 높은 변별력이 특정 분석 모형의 오차에 의한 것이 아니라, 실제 데이터의 패턴을 반영한 신뢰할 수 있는 결과임을 뒷받침한다.

셋째, 문항 반응의 안정성이다. 다음의 <그림4>와 같이, 모든 하위 유형에서 AI 문항은 인간 문항과 대등하거나 더 높은 변별도를 보였다. 이는 기존의 템플릿 기반 AIG가 갖고 있는 ‘단조로운 문항 생성’이나 ‘오답 논리 부재’의 한계를 극복했음을 보여주는 실증적 증거라고 할 수 있다. 즉, AI가 단순한 텍스트 생성을 넘어, 문항 반응 이론(IRT)이 요구하는 ‘변별력 있는 문항 특성 곡선(ICC)’을 구현할 잠재력을 갖추었음을 의미한다.

4.5 질적 연구: AI의 추론 과정과 학습자 인지 과정의 상호작용 분석

위와 같은 정량적 분석을 통해 AI가 생성한 문항이 인간 전문가의 문항과 심리측정적으로 동등하거나 더 우수한 변별력을 지니고 있음을 확인할 수 있었다. 그러나 그럼에도 불구하고 실제 학습자의 어떤 인지적 오류를 유발하여 변별력을 확보했는지는 구체적으로 설명하기 어렵다. 이에 본 절에서는 혼합 연구 방법론의 설명적 순차 설계(Explanatory Sequential Design)를 적용하여, AI 내부의 사고 과정 로그와 학습자의 사고 구술(Think-Aloud Protocol, TAP) 데이터를 통합적으로 분석했다. 이를 통해 문항 품질이라는 “결과” 뿐만 아니라 “과정”이라는 측면에서의 문항 타당도를 규명하고자 한다.

4.5.1 분석 대상 및 연구 절차

본 절의 질적 분석은 AI의 사고과정과의 통합 분석을 위해 <표5>와 같은 사고 과정을 통해 생성한 <표7>의 ‘유이어’ 문항(목표 어휘: ‘批判’)을 사례로 선정하여 진행했으며, 연구 절차는 다음의 2단계와 같다.

(4) 질적 분석의 연구 절차

- a. 사고 구술 데이터 수집 : 피험자(120명) 중 성적을 기준으로 선정한 12명(상·중·하 각 3명)을 대상으로 해당 문항(批判)을 푸는 과정에서의 모든 인지적 사고를 재현하도록 하는 동시적 사고 구술을 실시했다.
- b. 교차 검증 : <표5>의 AI가 예측한 ‘가상의 학습자 오류’와 실제 학습자의 ‘사고 구술’ 내용을 대조하여, AI 문항의 인지적 실재성을 검증한다.

4.5.2 학습자 사고 구술 분석 및 결과

AI가 의도한 설계 전략이 실제 학습 현장에서 유효하게 작동하는지 검증하기 위해, 상·중·하위 수준 학습자의 문제 해결 과정에 대한 분석을 수행했다. 총 12명의 피험자 중 성적을 기준으로 3명의 프로토콜을 발췌하여 미시적으로 분석한 결과는 다음의 (4)와 같다.

(4) 질적 분석의 연구 절차

- a. 상위권 학습자 A - 문맥적 단서(Contextual Clues)의 활용과 오답 소거
 - : 상위권 학습자 A는 상위권 학습자는 밑줄 친 批判이라는 단어 자체보다, 문장 전체의 흐름을 통해 해당 단어가 어떤 ‘의도’로 쓰였는지를 파악하는 데 집중했다.
 - (i) 사고 구술: “문제는 批判의 유이어를 찾는 거네요. 문장을 보면 무책임한 발언을 했고, 주위로부터 ‘厳しい’ 평가를 받았어요. 이걸 잘못을 지적받고 혼나는 상황이에요. (선택지를 보며) 1번 批評은 작품 같은 걸 분석하는 거니까 좀 딱딱하고 중립적이에요. 2번 非難... 이게 ‘나쁘게 말하다’니까 이 상황이란 딱 맞아요. 3번 평판은 문맥이 안 맞고... 정답은 2번.”
 - (ii) 분석: 피험자 A는 AI가 <표5>의 (2)문맥 설계에서 의도한 ‘부정적 문맥’을 정확히 포착했다. 批評을 ‘분석적·중립적’ 어휘로 분류하여 소거함으로써, AI가 설정한 의미적 함정을 성공적으로 회피했다.
- b. 중위권 학습자 B - 한자 형태와 L1 간섭의 혼동
 - : 중위권 학습자는 문맥보다는 제시된 단어의 ‘한자 모양’과 ‘한국어 번역’에 과도하게 의존하며 인지적 갈등을 겪었다.
 - (i) 사고 구술 : “批判이란 비슷한 말을 찾으려는 건데... 1번 批評이란 2번 非難 둘 다 말이 되는 것 같아요. 한국말로로는 ‘정부를 비평하다’, ‘정부를 비난하다’ 둘 다 쓰잖아요? (침묵 5초) 근데 批判의 ‘批’자랑 1번 批評의 ‘批’자가 똑같이 생겼어요. 한자가 같으니까 의미도 더 가깝지 않을까요? 2번 비난은 너무 감정적인 느낌이라... 그래서 1번 비평으로 했어요.”
 - (ii) 분석: 피험자 B는 AI가 오답 1로 설계한 함정에 정확히 걸려들었다. L1 간섭으로 인해 의미적 구분에 실패했고, 결국 확신이 없는 상태에서 ‘한자 형태의 유사성’이라는 표면적 단서에 의존하여 오답을 선택했다. 이는 AI가 로그에서 예측한 <표5>의 (3)오답설계의 ‘시각적 혼동’을 유발한다는 가설이 적중했음을 보여준다.

c. 하위권 학습자(Subject C): 하위권 학습자는 문장의 의미를 거의 파악하지 못하고, 선택지의 한자와 문제의 한자를 그림 맞추기 하듯 연결하려는 전략을 보였다.

(i) “무슨 뜻인지 잘 몰랐어요. 비판... 뒤에 글자가 ‘判’이네요. 보기 3번에 評判... 여기도 뒤에 ‘判’자가, 4번 判断도 ‘判’자가 있고... 보통 유의어 문제는 한자가 겹치는 게 답 아닌가? 3번 평판... 평판이 좋다 나쁘다 할 때 쓰는 말이니까... 그래서 3번으로 했습니다.”

(ii) 분석: 피험자 C는 문맥 단서를 활용하지 못하고, 표적 어휘 비판의 ‘判’자와 형태가 같은 글자가 들어간 評判을 정답으로 선택했다. 이는 AI가 오답 2를 설계할 때 ‘형태적 유사성’을 기반으로 설계한 바와 일치한다. 결과적으로 이 문항은 하위권 학습자를 변별해냈다.

4.5.3 통합 분석

AI의 ‘Deep Think’ 로그와 학습자의 TAP 데이터를 매핑하여 종합적으로 분석한 결과, AI가 생성한 문항이 단순한 ‘단어 찾기’가 아니라, 학습자의 어휘 이해 수준(심층적 이해 vs 표면적 이해)을 정확히 측정하는 도구로 작동하고 있음을 시사한다. 특히 批評과 評判이라는 오답지는 각각 중위권과 하위권을 걸러내는 강력한 필터 역할을 수행함으로써 문항의 전체적인 변별도를 높이는 데 기여했음을 확인할 수 있었다.

5. 결론

본 연구는 AI(Gemini 3.0 Pro)의 ‘Deep Think’를 활용하여 생성한 JLPT N2 어휘 문항이 인간 전문가의 문항과 심리측정적 동등성(TOST-P 검증)을 지니며, 오답 매력도와 변별도 측면에서는 오히려 더 우수함을 확인하였다. 단순한 텍스트 생성을 넘어 학습자의 인지적 오류까지 시뮬레이션하는 AI의 활용은, 문항 출제 시간을 약 90% 단축하고 수정률을 15% 미만으로 낮추는 압도적인 경제성을 입증했다. 이는 향후 컴퓨터 기반 적응형 검사(CAT) 도입 및 대규모 상시 평가 체제 전환을 위한 핵심적인 기술적 토대가 될 것이다.

또한, 이러한 기술적 진보는 개별 교사나 소규모 기관도 신뢰도 높은 ‘초개인화된 맞춤형 평가’ 도구를 자체 제작할 수 있게 함으로써, 교사의 역할을 ‘학습 설계자’로 재편하고 외국어 평가의 패러다임을 ‘선발’에서 ‘맞춤형 성장 지원’으로 견인할 것이다. 다만, 본 연구는 N2 레벨의 문자·어휘 영역과 제한된 전문가 패널 규모에 국한되었다는 한계가 있으므로, 향후 독해·청해 및 최상위 레벨(N1)로의 확장과 대규모 표본을 통한 후속 연구가 요구된다.

◀ 참고문헌 ▶

김유영(2025) 「페르소나 기반 로컬 LLM 챗봇과 AI 프롬프트를 활용한 일본어 교육 프레임워크—비판적 AI 리터러시와 학습자 주도성 함양을 중심으로—」 『일본문화학보』 107집, 한국일본문화학회, pp.303-312. DOI: 10.21481/jbunka..107.202511.303

国際交流基金他(2009) 『新しい「日本語能力試験」ガイドブック 概要版』, 国際交流基金, pp.1-2

- Ayanwale, M., Chere-Masopha, J., & Morena, M. C.(2022) “The classical test or item response measurement theory: the status of the framework at the examination council of Lesotho”, *International Journal of Learning, Teaching and Educational Research*, 21(8), Society for Research and Knowledge Management Ltd, pp.384-406. DOI:doi.org/10.26803/ijlter.21.8.22
- Baker, T., Smith, L., & Anissa, N.(2019) “Educ-AI-tion rebooted? Exploring the future of artificial intelligence in schools and colleges”, Nesta Foundation, pp.1-56. https://media.nesta.org.uk/documents/Future_of_AI_and_education_v5_WEB.pdf
- Leo M. Harvill.(1991) “Standard Error of Measurement”, An NCME Instructional Module on, Educational Measurement, 10(2), National Council on Measurement in Education, pp.33-41. <https://doi.org/10.1111/j.1745-3992.1991.tb00195.x>Digital Object Identifier (DOI)
- Daniël Lakens & Anne M. Scheel(2018) “Equivalence Testing for Psychological Research: A Tutorial”, *Advances in Methods and Practices in Psychological Science*, 1(2), Association for Psychological Science, pp.259-269. <https://doi.org/10.1177/2515245918770963>
- Lord, F.M.(1952) “The relationship of the reliability of multiple-choice test to the distribution of item difficulties. *Psychometrika*, 18, pp.181-194. <https://doi.org/10.1007/BF02288781>
- Malau-Aduli, B.S., Walls, J., & Zimitat, C.(2012) “Validity, reliability and equivalence of parallel examinations in a university setting”, *Creative Education*, 3, Scientific Research Publishing, pp.923-930. DOI: 10.4236/ce.2012.326140
- Mara, C.A., & Cribbie, R.A.(2012) “Paired-samples tests of equivalence”, *Communications in Statistics - Simulation and Computation*, 41(10), Taylor & Francis Group, pp.1928-1943. DOI:doi.org/10.1080/03610918.2011.626545
- Mark J Gierl, Hollis Lai, Simon R Turner(2012) “Using automatic item generation to create multiple-choice test items”, *Medical Education*, 46(8), The Association for the Study of Medical Education, pp.757-765. DOI: 10.1111/j.1365-2923.2012.04289.x
- O, K.-M.(2024) “A comparative study of AI-human-made and human-made test forms for a university TESOL theory course”, *Language Testing in Asia*, Asian Association for Language Assessment, pp.14-19. <https://doi.org/10.1186/s40468-024-00291-3>
- D. Precht, C. Hazlett, S. Yip, J. Nicholls & Y. Jin(2003) “Item analysis user’s guide. International Database for Enhanced Assessments and Learning”, The IDEAL Consortium, pp.1-69
- Schuirmann, D.J.(1987) “A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability”, *Journal of Pharmacokinetics and Biopharmaceutics*, 15, Springer Science+Business Media, pp.657-680. DOI: 10.1007/BF01068419
- Shin, D., & Lee, J.H.(2023) “Can ChatGPT make reading comprehension testing items on par with human experts?”, *Language Learning & Technology*, 27(3), National Foreign Language Resource Center, pp.27-40. <https://doi.org/10.64152/10125/73530>
- ioJLPT, ioJLPT 온라인 모의 일본어능력시험 www.onlinejplt.com (2026년 1월 10일 검색)

〈Abstract〉

**Beyond Generation to Reasoning: Psychometric Equivalence of JLPT Items
Developed via Gemini 3.0 Pro's "Deep Think" Mode**

This study empirically investigates the psychometric equivalence of Japanese Language Proficiency Test (JLPT) N2 vocabulary items generated via Gemini 3.0 Pro's "Deep Think" capability compared with those developed by human experts. Existing research on large language model (LLM)-based item generation has been predominantly Anglocentric, often failing to address the complex nuances and distractor design logic required for high-context languages such as Japanese. To address this limitation, we propose a "Reasoning-Enhanced Agentic Workflow" that transcends simple text generation by emulating the cognitive processes of human examiners to infer and validate learner error patterns. A paired-sample study involving 120 Japanese learners utilized the TOST-P method, confirming that AI-generated items achieved statistical equivalence to human-generated benchmarks. Significantly, item response analysis revealed that the AI items exhibited superior discrimination indices and distractor attractiveness. These findings indicate that the model can effectively simulate phonological and morphological misconceptions to engineer functional distractors. Consequently, this study establishes the value of AI not merely as an auxiliary tool but as a primary "educational agent" capable of autonomously architecting high-quality assessment instruments.

Key Words : AI, automated item generation (AIG), psychometric equivalence, TOST-P, JLPT

◆ 金晳泳(Kim, Yu-Young)

- 소속 : 동덕여자대학교 일어일본학과 부교수
- E-mail : yuiyu@hotmail.com

논문투고일	2026. 01. 31.
심사개시일	2026. 02. 05.
심사완료일	2026. 02. 21.
게재확정	2026. 02. 25.