

생애사 프로필 기반 LLM 소설 생성의 창의성 평가 연구: 모델별 서사 생성 특성 비교를 중심으로

이 만 유¹ · 김 예 진² · 전 승 화³ · 오 동 석^{4*}

¹경북대학교 영어영문학과 석박사통합과정

²조지워싱턴대학교 컴퓨터공학과 박사과정

³경북대학교 불어불문학과 & 인문카운슬링학과 부교수

⁴경북대학교 영어영문학과 조교수

Evaluating the Creativity of Life-History-Based LLM Story Generation: A Comparative Analysis of Model Narrative Features

Man-Yu Lee¹ · Yejin Kim² · Seung-Hwa Jean³ · Dong-Suk Oh^{4*}

¹Integrated Master-Doctoral Degree Programs, Department of English Language and Literature, Kyungpook National University, Daegu 41566, Korea

²Ph.D Course, Department of Computer Engineering, George Washington University, Washington, DC, USA

³Associate Professor, Department of French Language and Literature & Department of Humanistic Counseling, Kyungpook National University, Daegu 41566, Korea

⁴Assistant Professor, Department of English Language and Literature, Kyungpook National University, Daegu 41566, Korea

[요 약]

본 연구는 대규모 언어모델(Large Language Models; LLMs)의 생애사 소설 생성 가능성을 탐구하고, 창의성 평가 기준에 따른 인간 평가를 통해 그 성과를 검증하였다. 노년층 심층 인터뷰를 기반으로 구성된 내러티브 프롬프트를 활용하여 GPT-3.5, GPT-4o, GPT-4.1 모델이 각각 100편씩 총 300편의 생애사 소설을 생성하였으며, 여섯 가지 창의성 평가 기준에 따라 인간 평가자가 정량·정성 분석을 수행하였다. 분석 결과, GPT-3.5는 단순한 구조로 정교성에서 가장 높은 합의를 보였고, GPT-4o는 안정적인 서사 전개를 통해 전반적으로 균형 잡힌 성과를 나타냈다. GPT-4.1은 감각적·정서적 표현에서 강점을 보였으나 해석이 다양하게 분기되어 항목별 일치도 편차가 가장 크게 나타났다. 이는 창의성 수준의 차이보다는 각 모델이 구성하는 서사 방식이 평가자 해석에 다르게 작용한 결과이며, LLM이 생애사와 같은 복잡한 맥락에서도 일정 수준의 창의적 서사 생성 능력을 발휘함을 시사한다.

[Abstract]

This study examines the ability of large language models (LLMs) to generate creative life-history narratives. Using prompts derived from interviews with older adults, GPT-3.5, GPT-4o, and GPT-4.1 produced 100 stories each, which were evaluated by human raters across six creativity criteria. GPT-3.5 showed high agreement in elaboration owing to its simple structure, GPT-4o demonstrated stable and balanced narrative development, and GPT-4.1 excelled in sensory expression but yielded wider variations in rater interpretations. These results indicate that differences in agreement arise from how each model constructs its narratives, and that LLMs can produce creative narratives even in contextually complex domains such as life-history storytelling.

색인어 : 대규모 언어모델, 서사 생성, 생애사, 창의성 평가, 개인 맞춤형 스토리텔링

Keyword : Large Language Model, Narrative Generation, Life History, Creative Evaluation, Personalized Storytelling

<http://dx.doi.org/10.9728/dcs.2026.27.1.13>



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 30 October 2025; Revised 01 December 2025

Accepted 02 January 2026

*Corresponding Author; Dong-Suk Oh

Tel: +82-53-950-7155

E-mail: inow3555@knu.ac.kr

1. 서론

최근 트랜스포머(Transformer) 아키텍처의 도입으로 시작된 언어모델의 혁신은 대규모 언어모델(Large Language Models, LLM)이 단순한 언어 처리를 넘어 창작 활동에 참여하는 수준에 이르게 하였다[1]. ChatGPT를 비롯한 LLM 연구에서 창의성 평가는 중요한 우선순위로 지적되고 있으며[2], 특히 스토리텔링 영역에서 LLM의 장편 서사 생성 능력이 주목받고 있다[3]. 그러나 지금까지의 연구는 주로 단편적 창작물이나 짧은 서사 수준에서의 평가에 집중되어 왔으며[4], LLM이 장기적 맥락을 유지하며 의미 있는 서사를 구성할 수 있는가에 대한 체계적 평가는 부족하다. 이에 본 연구는 ‘생애사 소설(Life history novel)’이라는 복합적 장르를 대상으로 LLM 창작물의 창의성 평가 체계를 검증하고자 한다. 생애사 소설은 개인의 정체성, 시간적 인과성, 사회문화적 맥락을 장기간에 걸쳐 일관되게 유지해야 하는 복잡한 서사 구조를 지니므로, LLM의 창의적 서사 능력을 평가하기에 적합하다.

창의성 평가의 이론적 기반은 Torrance의 고전적 4요소 모델(유창성, 유연성, 독창성, 정교성)에서 출발하였다[5]. 이후 Shively 등은 이를 확장하여 유용성(utility)과 구체적인 창의성 전략(Specific creativity strategy)을 추가로 제안하였으며[6], Kim&Oh는 이를 바탕으로 총 6요소를 정식화된 LLM이 창작한 글을 평가하였다[7]. 특히 이 연구에서는 유창성과 정교성 차원에서 높은 평가자 간 일치도(Inter-Annotator Agreement; IAA)를 보고하였으나, 이러한 높은 IAA가 실제 창의성의 향상인지 평가 용이성에서 기인한 것인지는 추가 검증이 필요하다. 최근에는 LLM의 산출물이 인간 평가자의 창의성 인식에 어떠한 영향을 미치는지 탐구하는 연구들이 등장하고 있으나[8], 장편 서사와 같이 복합적인 서사(Narrative)구조를 포함한 평가 연구는 여전히 드물다.

생애사 연구는 개인의 삶을 단순히 회고하는 기술이 아니라, 삶의 궤적 속에서 경험한 사건과 선택의 의미를 이해하려는 질적 탐구 과정으로 발전해 왔다[9]. 특히 노년기 연구에서는 과거의 생애 경험이 현재 삶의 질과 정체성 형성에 미치는 영향을 분석하는 데 유용한 방법으로 주목받고 있으며, 개인이 격동의 사회·역사적 맥락 속에서 어떻게 삶을 재구성해 왔는지를 탐색한다. 또한 최근에는 기록학의 관점에서 생애사 연구를 개인 기록화의 이론적 근거로 확장하려는 시도가 이루어지고 있다[10]. 이러한 접근은 개인의 기록물이 담고 있는 특수성과 보편성을 통해 개인과 사회의 관계를 해석하고, 개인의 가치와 경험을 사회적 맥락 속에서 지속적으로 보존하려는 학문적 방향을 제시한다.

이러한 접근은 개인의 목소리와 의미 구성을 섬세하게 드러내는 장점을 가지지만, 동시에 시간과 자원의 제약, 연구자 편향, 대규모 비교 연구의 어려움이라는 한계를 지닌다. 본 연구는 이러한 생애사 연구 전통을 LLM 창작 평가와 결합하여,

6요소 창의성 평가 틀을 서사에 적용한다는 점에서 방법론적 독창성을 지닌다.

연구 설계를 위해 OpenAI의 GPT-3.5, GPT-4, GPT-4o 세 모델을 사용하여 총 300편의 생애사 소설을 생성하였다. 프롬프트는 인물의 기본 정보(연령, 성별, 직업, 거주지)와 주요 생애 사건(교육, 결혼, 직업 경력, 전환점)을 포함하도록 구조화하였으며, 각 모델당 100편씩 균등하게 생성하였다. 이렇게 수집된 텍스트는 사전 훈련된 두 명의 평가자가 6요소 창의성 평가 체계에 따라 독립적으로 채점하였다.

본 연구는 다음 세 가지 핵심 질문에 답하고자 한다: (1) 6요소 평가 체계가 LLM이 생성한 생애사 소설 평가에 적합한가? (2) 생애사 소설이라는 복잡한 장르에서도 높은 평가자 간 일치도가 유지되는가? (3) 서로 다른 LLM 모델 간 생애사 소설의 창의성 평가에서 서사 생성 능력에 유의미한 차이가 존재하는가?

본 연구는 생애사 서사를 LLM의 창의적 서사 생성 능력 평가를 위한 실험적 장르로 선정하였다. 생애사 서사는 개인의 정체성, 시간적 인과성, 사회문화적 맥락이 복잡하게 얽힌 장기적 내러티브 구조를 요구한다는 점에서, LLM이 인간의 창의적 사고와 서사적 정체성 구성 능력을 어느 정도 재현할 수 있는지를 검증하기에 적합한 도메인이다. 본 연구에서 수행한 LLM 소설 생성물의 창의성 평가 과정은 그림 1과 같다. 그림 1에서 보이는 바와 같이, 본 연구는 생애사 데이터를 기반으로 프롬프트를 구성하고, 세 가지 LLM 모델(GPT-3.5, GPT-4o, GPT-4.1)을 통해 소설을 생성한 후, 두 명의 평가자가 6가지 창의성 기준(유창성, 유연성, 독창성, 정교성, 유용성, 구체적인 창의성 전략)에 따라 평가하는 과정을 거친다.

본 연구의 핵심 기여는 6요소 창의성 평가 틀을 생애사 소설이라는 서사에 적용하여 LLM이 생애사와 같이 장기적 맥락과 정체성 변화를 보이는 서사에서 창의적 평가 능력이 적용될 수 있음을 실증적으로 보여주는 데 있다. 생애사 소설에서의 창의성 평가는 기존의 단편 창작물 평가와 근본적으로 다른 도전을 제시한다. 유창성은 단순한 언어적 유려함을 넘어 장기간의 서사적 일관성으로, 정교성은 감정 곡선과 정체성 형성의 진정성으로 확장된다. 독창성과 유연성은 개인 경험의 고유성과 문화적 맥락의 적절한 반영을 통해 평가되며, 유용성과 구체적인 창의성 전략은 서사의 진정성과 문화적 의미 구성 능력을 중심으로 검증된다. 이러한 확장은 단순한 평가 틀의 적용을 넘어, 창의성 개념을 서사적 차원에서 재해석하고자 하는 시도이다.

이 연구는 단순히 평가 기준의 타당성을 검증하는 데 그치지 않고, LLM이 생성한 서사의 특성이 평가자 간 일치도에 어떠한 영향을 미치는지를 탐구한다. 기존 연구들은 특정 항목에서 높은 일치도를 제시했지만, 이러한 결과가 모든 유형의 서사에 일반화될 수 있는지는 검증되지 않았다. 본 연구는 생애사 소설이라는 복합적 서사를 통해 동일한 평가 기준 하에서도 평가자에게 서로 다른 해석 지점을 유도하며, 이것이 항목별 일치도의 차이로 이어짐을 실증적으로 분석한다.

나아가 본 연구는 LLM 창의성 평가에서 '높은 창의성 점수'와 '높은 평가자 간 일치도'가 반드시 일치하지 않는다는 점을 드러내어, 향후 LLM 창작물 평가 연구에서 평가 기준의 정교화와 서사 유형별 맞춤형 평가 체계 개발의 필요성을 제시하고자 한다. 이를 통해 생애사라는 특수 도메인에서 LLM의 창의적 서사 생성 가능성과 한계를 종합적으로 조명하고, 개인 기록화 및 생애사 연구 분야에서 LLM 활용의 실질적 기반을 마련하는 데 기여하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 LLM의 기술적 발전 과정과 생애사 연구의 이론적 배경을 검토하고, LLM 창의성 평가에 관한 연구를 분석한다. 3장에서는 연구 설계로서 생애사 데이터 수집과 구조화, 프롬프트 설계, LLM을 통한 소설 생성 과정, 그리고 6요소 창의성 평가 체계와 평가 절차를 상세히 기술한다. 4장에서는 창의성 항목별 정량 분석 결과와 모델별 서사 특성을 비교하고, 실제 생애사와 LLM 생성 서사의 질적 차이를 분석한다. 마지막으로 5장에서는 연구 결과를 종합하여 의의와 한계를 논의하고 향후 연구 방향을 제시한다.

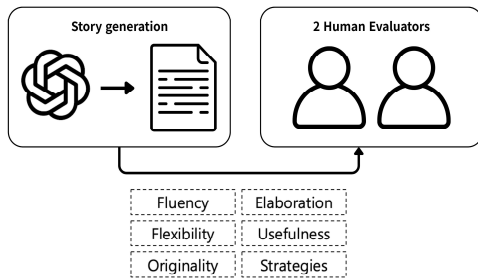


그림 1. LLM 소설 생성물의 창의성 평가 과정
Fig. 1. The process of evaluating LLM-generated novels for creativity

II. 대규모 언어모델 기반 생애사 서사와 창의성 평가 이론

2-1 대규모 언어모델의 발전

대규모 언어모델(Large Language Models; LLMs)은 트랜스포머(Transformer) 아키텍처의 도입 이후, 자연어 이해와 생성의 방식을 근본적으로 바꾸었다. 트랜스포머는 순차적 구조를 제거하고 어텐션(attention) 메커니즘을 통해 단어 간 관계를 병렬적으로 학습함으로써 장기 의존성을 효과적으로 포착할 수 있게 하였다. 이러한 구조적 혁신은 이후 수많은 인코더-디코더 계열 모델의 발전을 촉발시켰다.

인코더 기반 모델은 문맥 이해 능력에 초점을 맞추었다. BERT(Bidirectional Encoder Representations from Transformers)는 마스크드 언어모델(Masked Language Modeling; MLM)과 문장 예측(Next Sentence Prediction;

NSP) 과제를 결합하여, 양방향 문맥 정보를 동시에 학습할 수 있도록 설계되었다[11]. 이후 RoBERTa는 데이터 크기와 학습 전략을 조정해 BERT의 일반화 성능을 개선했으며, ALBERT는 파라미터 공유와 문장 순서 예측(Sentence Order Prediction; SOP)을 통해 모델 효율성을 크게 높였다[12].

OpenAI의 GPT-1은 비지도 사전학습(pre-training)과 지도 미세조정(fine-tuning)을 결합하여 자연어 생성에 필요한 일반적 언어 패턴을 학습할 수 있음을 최초로 입증하였다[13]. 약 1억 개의 파라미터를 기반으로 한 GPT-1은 모델이 데이터셋의 통계적 특성을 학습함으로써, 제한된 지도 데이터만으로도 다양한 언어 과제에 적용될 수 있음을 보여주었다. 이는 LLM의 학습 패러다임이 대규모 사전학습에서 소규모 미세조정 구조로 전환되는 계기가 되었다.

GPT-2는 GPT-1의 구조를 확장하여 15억 개 이상의 파라미터를 학습시킴으로써, 더욱 장기적인 문맥을 유지하고 일관된 텍스트를 생성할 수 있음을 보여주었다[14]. 특히 인간이 작성한 텍스트와 거의 구분되지 않는 수준의 언어 품질을 보이며, 대중적 주목을 받았다. 이 시기에는 모델 크기와 성능 간의 관계를 수학적으로 규명한 Scaling Laws 연구가 등장하여, 파라미터 수·데이터 크기·연산량이 성능 향상과 선형적으로 연관된다는 사실을 제시하였다[15]. 이러한 결과는 이후 LLM 개발의 규모 확장 전략을 뒷받침하는 이론적 근거가 되었다.

GPT-3은 1,750억 개의 파라미터를 학습하며, LLM이 특정 과제에 특화된 모델이 아니라 범용적 언어 생성 기능을 수행할 수 있음을 입증하였다[16]. 특히 GPT-3은 별도의 미세조정 없이도 zero-shot, one-shot, few-shot 학습을 수행할 수 있는 능력을 보여주었으며, 이는 “예시만으로 새로운 과제를 해결하는” 언어모델의 추론 능력을 처음으로 실증한 사례였다. 이로써 GPT 계열은 단순한 언어모델을 넘어, 인지적 추론과 생성 능력을 통합한 범용 인공지능으로 발전하게 되었다.

가장 최근의 GPT-4는 텍스트 입력뿐 아니라 이미지·음성 등 멀티모달 입력을 통합적으로 처리할 수 있는 기능을 갖추었다[17]. 이를 통해 언어 이해와 시각적 추론이 결합된 복합 작업이 가능해졌으며, 대화 응답의 맥락 유지와 사실성도 크게 향상되었다. 최신 버전인 GPT-4o는 오디오·비디오 입력을 포함한 실시간 상호작용형 모델로 진화하여, 언어적 창의성뿐 아니라 감정적·서사적 요소까지 포착할 수 있는 환경을 제공한다[18].

이처럼 LLM의 발전은 단순한 파라미터 확장이나 성능 향상을 넘어, 맥락 유지 능력과 서사적 일관성을 강화하는 방향으로 진화해왔다. 이러한 기술적 진보는 본 연구에서 다루는 생애사 소설 생성과 같은 장기적 서사 텍스트를 분석하고 평가하는 데 필수적인 기반을 제공한다. 다시 말해, LLM의 구조적·기능적 발전은 창의적 언어 생성 능력 평가의 기술적 전제 조건을 마련했다고 할 수 있다.

2-2 대규모 언어모델에서의 생애사 서사

생애사(Life History Research) 연구는 개인을 대상으로 면담을 통해 수집한 구술자료를 해석한 결과를 삶의 이야기로 기술하는 데 초점을 두고 있다. 이런 측면에서 생애사는 자서전, 전기, 사례사, 구술사 등의 내러티브 연구보다 한걸음 더 나아간 연구방법이다. 개인이 살아온 삶의 경험은 단편적 사실로 온전히 포착할 수 없다. 이런 삶의 이야기를 온전히 이해하기 위해 주체와 객체, 개인과 사회, 행위와 구조의 경계를 허물어, 삶의 이야기가 품고 있는 배경, 즉 시간과 장소의 의미를 탐색하는 생애사 연구가 필요하다. 이때, 생애사 연구에서 삶의 이야기는 맥락을 발견하는 창문의 구실을 한다 [19]. 이러한 서사적 접근은 인간의 정체성, 감정, 관계의 변화를 통시적으로 이해하고, 개인 경험이 사회적 구조와 상호 작용하며 창의적으로 재구성되는 과정을 보여준다.

또한, 생애사 연구의 내러티브적 성격을 ‘Narradigm’이라 명명하며, 인간의 삶과 지식이 본질적으로 서사적 구조 안에서 인식되고 재현된다고 주장되었다[20]. 이 연구에서 생애사 연구는 전통적 실증주의의 단일한 진리 개념을 넘어 개인의 경험을 다층적·맥락적으로 해석하는 방법론적 다원주의를 제안한다고 보았다. 즉, 삶의 이야기를 연구 대상으로 삼는 것은 단순한 자서전적 기술이 아니라, 개인적 기억과 사회적 담론이 교차하는 인식론적 장을 탐색하는 일이다. 이러한 관점은 인간이 언어를 통해 자신과 세계를 해석한다는 점에서, 오늘날 LLM이 수행하는 언어 생성 행위와 근본적으로 맞닿아 있다.

최근 LLM은 맥락 유지, 감정 표현, 사건 간 인과 구조의 재현 능력을 통해 인간의 서사적 사고를 일정 부분 모사하고 있다. 특히 GPT-4 계열 모델은 인물의 성장, 상실, 회복, 전환점 등 장기적 맥락을 포착하여 이야기의 구조를 유지할 수 있는 능력을 보인다. 이러한 특성은 생애사 연구가 강조해 온 시간성과 맥락적 의미 구성의 원리와도 일치한다. 그러나 LLM의 서사 생성이 인간의 창의적 서사와 동일한 수준의 깊이와 통찰을 지니는지는 여전히 논쟁적이다. 최근 연구에서는 LLM의 스토리 생성 능력을 자동 평가 지표와 인간 평가를 병행해 분석한 결과, 기존 스토리 생성 모델보다 높은 유창성과 구조적 완결성을 보였으나 세계 지식의 불일치, 서사적 참신성 부족, 표절 유사성 등의 한계가 확인되었다[21].

따라서 생애사 서사는 LLM의 창의적 서사 생성 능력을 평가하기 위한 적절한 실험적 장르로 간주될 수 있다. 다시 말해, LLM이 생애사 서사를 얼마나 일관되고 의미 있게 구성하는지를 분석하는 것은 곧 모델이 인간의 창의적 사고와 서사적 정체성 구성 능력을 어느 정도 재현할 수 있는가를 검증하는 과정과 직결된다.

2-3 대규모 언어모델에서의 창의성 평가 연구

LLM이 단순한 문장 생성기를 넘어 스토리텔링, 시, 에세이 등 창작적 언어 활동에 활용되면서, 산출물의 창의성

(creativity)을 평가하려는 연구가 빠르게 확산되고 있다. 최근 연구에 따르면, 생성형 AI(Generative AI)는 일부 창의성 테스트에서 인간과 유사한 수준의 성능을 보이며, 전 세계 지식 근로자의 창의적 역량을 증진시키는 잠재력을 지닌 것으로 나타났다[22]. 이러한 결과는 LLM이 단순히 언어적 반복을 수행하는 모델이 아니라, 인간의 창의적 사고 과정 일부를 모방하고 강화할 수 있음을 시사한다. Keskler 등은 CTRL(Control Transformer) 모델을 통해 문체, 주제, 정서 등을 조절하며, 언어모델의 생성 과정 자체가 창의적 조합을 수행할 수 있음을 입증했다[23]. 이러한 제어 생성(controlled generation) 연구는 모델이 단순히 데이터를 복제하는 것이 아니라, 문맥적 변형과 의미 확장을 통해 의도적 창작 행위를 수행할 수 있음을 보여준다.

Zhao 등은 LLM 산출물의 창의성을 자동 지표와 인간 평가를 병행하여 분석했으며, 유창성(fluency)·정교성(elaboration)·유연성(flexibility) 차원에서는 높은 평가를 받았으나, 독창성(originality)과 진정성(authenticity) 측면에서는 여전히 한계가 있음을 보고하였다. Kim&Oh[7]는 LLM이 생성한 글 및 인간이 쓴 글을 평가용 6요소 체계(유창성, 유연성, 독창성, 정교성, 유용성, 구체적인 창의성 전략)를 기준으로 AI(Artificial Intelligence) 평가를 하였고, 이를 평가자 간 일치도(IAA)로 분석하였다. 그 결과, 언어적 요소에서는 높은 일관성을 보였지만, 해석적 요소에서는 평가자 간 차이가 두드러졌다.

또한 최근 연구들은 LLM을 평가자로 활용하는 가능성도 탐색하고 있다. Li 등은 LLM이 인간 평가자의 기준을 어느 정도 재현할 수 있지만, 세부 항목 간 신뢰도 편차가 크다는 점을 지적하였다[24]. Yi 등은 서사의 일관성을 평가하기 위한 SCORE(Story Coherence and Retrieval Enhancement) 모델을 제안하여, 자동화된 내러티브 평가 지표의 가능성을 확장하였다[25]. 그러나 이러한 접근은 여전히 짧은 서사나 단일 주제 중심의 평가에 머물러 있으며, 생애사 소설처럼 시간적 인과성, 정체성, 사회문화적 맥락이 복합적으로 얽힌 장르에 대한 체계적 평가는 이루어지지 않았다.

III. 대규모 언어모델 기반 생애사 소설 생성 및 창의성 평가 방법

3-1 데이터 설계 및 소설 데이터 생성

1) 연구 대상

본 연구에서 활용된 데이터는 생애사 연구팀이 직접 수행한 심층 면담을 통해 수집된 고령층 대상 생애사 기록에 기반한다. 총 10명의 참여자가 모집되었으며, 모집 과정에서는 참여자의 연령(65세 이상), 기본 의사소통 능력, 연구 동의 여부를 고려하였다. 연구 참여 전 모든 대상자는 연구 목적과 활용 방식에 대해 설명을 듣고 서면 동의서를 제출하였다.

면담은 일회성 대화가 아닌 수차례의 만남으로 이루어졌으

며, 응답자의 연령대(65~81세), 건강 상태, 기억력 차이, 정서적 반응 등 다양한 변인이 개입되었다. 동일한 주제를 다루더라도 이야기 전개 방식, 감정의 밀도, 표현 양식은 크게 달라, 이를 일정한 텍스트 형식으로 통일하는 과정에서 많은 어려움이 뒤따랐다. 특히 서사 단위의 변동성을 줄이기 위해, 연구팀은 반복 청취와 교차 검증 과정을 거쳐 발화 내용을 구조화된 텍스트로 정리하였다.

본 연구의 소규모 질적 표본은 이러한 전체 집단의 대표성을 갖추고 있지는 않지만, 실제 현장에서의 구술 변동성과 서사적 다양성을 반영하는 데 초점을 두었다. 따라서 이 10명의 생애사 기록은 자료 수집 과정의 복잡성과 해석의 난이도를 내포하고 있으며, 이는 LLM 기반 생애사 서사 자동화 연구의 맥락에서 중요한 출발점이 된다.

이질적인 면담 데이터를 LLM이 활용 가능한 학습 및 생성 자료로 전환하기 위해, 연구팀은 면담 내용을 서사 구성의 핵심 요소에 해당하는 항목으로 분류·체계화하였다. 초기에는 기본 정보(Basic), 개인적 주체성(Agency), 가족 및 관계성(Communion), 사회적 맥락(Society), 전환점(Turning Point), 이야기의 일관성(Coherence) 등 여섯 가지 항목으로 분류하였다[26],[27]. 그러나 예비 분석(pilot analysis)에서 일부 항목 간 중복성과 경계 모호성이 확인되었으며, 특히 Basic과 Coherence는 다른 항목의 맥락 속에서 충분히 포착 가능하다는 결론에 도달하였다. 이에 따라 연구팀은 전문가 협의와 주제 분석을 거쳐 최종적으로 Agency, Communion, Turning Point의 세 가지 항목으로 통합·단순화하였다.

이 과정에서, 예를 들어 Basic은 Agency 내 개별적 배경 서술에 통합되었고, Coherence는 Turning Point와의 연계성을 통해 유지되었다. 또한 Communion과 Society는 상호 중첩적 요소가 많아 'Communion'으로 통합하였다. 이렇게 축소된 체계는 항목 수를 줄여 효율성을 확보하는 동시에, 주요 내러티브 의미망을 보존하여 서사적 심층성을 손상시키지 않도록 질적 내용분석 원칙을 적용하였다.

- (1) Agency: 개인의 주체적 특성과 삶의 태도, 정체성 형성과 같은 내적 요인.
- (2) Communion: 가족 관계, 사회적 교류, 직업적 경험, 사회 변화에의 적응 등 개인이 속한 관계망과 환경적 요인.
- (3) Turning Point (TP): 개인의 삶에서 중요한 변화를 이끄는 사건, 위기 극복, 새로운 의미 형성의 순간.

이러한 3가지 항목은 생애사 서사의 핵심 구성 요소로, LLM이 생애사적 내러티브를 생성할 때 반드시 반영해야 하는 구조적 기반을 제공한다. 즉, Agency는 주체성·개인적 특성을, Communion은 사회적 맥락·관계적 요인을, Turning Point는 삶의 전환과 서사의 긴장·해소 구조를 나타내며, 이 세 요소가 균형 있게 반영될 때 설득력 있는 생애사 소설이 완성된다.

연구팀은 면담 기록을 단순 요약하는 데 그치지 않고, 각

항목별로 시간적 흐름과 맥락적 의미를 보존하면서도 LLM 프롬프트 입력으로 활용 가능한 구조적 단위로 정제하였다. 결과적으로 이 데이터 구조화 작업은 (1) 프롬프트 설계의 핵심 기반이 되었으며, (2) 다양한 생성 결과를 비교·분석할 수 있는 기준틀로 기능하였다.

표 1은 실제 면담 자료를 Agency, Communion, Turning Point 항목으로 정리한 입력 데이터 예시를 제시하며, LLM 프롬프트 입력의 구조적 기반을 시각적으로 보여준다.

표 1. 생애사 서사 항목 3가지와 각 항목별 데이터 구성 예시

Table 1. Three life history narrative categories and example data for each category

Dataset	Contents
Agency	<ul style="list-style-type: none"> • School Memory: "When I was in elementary school, we had the 'Saemaedul Movement', so I would wake up early in the morning to clean the neighborhood with a broom. It was fun and became a driving force in developing my sense of independence and life skills."
Communion	<ul style="list-style-type: none"> • Deep Affection for Mother: "My father passed away early, and we became financially unstable. My mother had to work alone to support us. During my first winter break in middle school, I suddenly got severely ill due to internal inflammation and couldn't attend school for several months." • Social Issues in life: "During my military service, the Aung San terrorist bombing occurred, putting our unit on high alert."
Turning point	<ul style="list-style-type: none"> • Major turning point in life: "After the early death of my father, who was a postal delivery worker, my mother became heavily dependent on me emotionally. She overcame this through religion, and I also wish to reflect this in the output of my life story."

2) 프롬프트

본 연구에서 활용된 프롬프트는 단순히 결과를 요청하는 방식이 아니라, 모델이 사고 과정을 차근차근 전개하도록 유도하는 구조적 설계를 적용하였다. 이는 복잡한 문제 해결이나 장기적 맥락 유지가 요구되는 작업에서 모델의 성능을 향상시킬 수 있는 기법으로 알려져 있으며, 본 연구에서는 생애사 소설이라는 다층적 내러티브를 구성하는 데 활용되었다.

구체적으로, 프롬프트는 "인물의 전기적 배경 → 삶의 전환점 → 감정 곡선 → 사회적 관계"와 같은 서사의 핵심 단계를 순차적으로 제시하였다. 단순히 "생애사 소설을 작성하라"는 지시 대신, 모델이 각 단계에서 고려해야 할 요소를 차례로 안내함으로써 추론 과정이 단계적으로 전개되도록 설계한 것이다. 예를 들어, 인물의 전기적 배경을 먼저 설명하고, 이어서 삶에서 중요한 전환점을 서술하게 한 후, 그 전환점이 인물의 정체성과 감정 곡선에 어떤 영향을 미쳤는지를 풀어내도록 유도하였다. 마지막으로 사회적 관계와 상호작용을 통해 내러티브가 완결되도록 구조화하였다.

이와 같은 접근은 복잡한 수학 문제나 다단계 추론이 필요

한 질의응답 상황에서 모델이 “생각을 말로 풀어내며 답을 찾는 방식”으로 성능을 향상시키는 사례에서 착안하였다[28]. 본 연구에서는 동일한 원리를 서사 생성에 적용하여, 모델이 삶의 사건을 시간적·논리적으로 연결하고 감정적 흐름을 정교하게 반영할 수 있도록 했다.

결과적으로, 단계적 추론 과정을 내재화한 프롬프트는 모델이 산출하는 텍스트의 맥락적 일관성, 감정적 설득력, 사회적 관계망 재현을 강화하는 효과를 가져왔다. 이는 단순 프롬프트와 비교했을 때, LLM이 더 정교하고 창의적인 생애사 소설을 생성할 수 있는 기반이 되었다. 따라서 표 2는 소설 생성을 위한 실제로 설계한 프롬프트 예시이다.

3) 소설 데이터

생애사 소설 생성에는 세 가지 최신 LLM 모델이 활용되었다. 구체적으로 GPT-3.5, GPT-4o, GPT-4.1 모델을 선정하였으며, GPT-3.5는 비교 기준(baseline)의 역할을 담당하고, GPT-4o와 GPT-4.1은 최신 모델로서 창의성 발현 수준을 검증하기 위해 포함하였다. 본 논문에서 사용된 300편의 소설은 OpenAI API(Application Programming Interface)를 활용한 자동화된 스크립트가 동일한 조건에서 일괄적으로 생성한 결과물이다.

각 소설 생성은 완전히 독립된 대화 세션에서 수행되었으며, 이전 생성물이나 대화 기록이 후속 생성에 어떠한 형태로도 반영되지 않도록 매 샘플마다 세션을 초기화하였다. 이는 최신 LLM이 long-context 정보를 적극적으로 활용하는 특성을 고려하여, 모델 간 비교에서 발생할 수 있는 공정성 문제를 방지하고 통계적으로 독립된 표본을 확보하기 위한 조치이다. 이 동일한 절차는 모든 모델에 일관되게 적용되었다.

각 모델에 대해 동일한 인물 프로필과 동일한 프롬프트 구조를 적용하여 모델당 100편씩 총 300편의 생애사 소설을 연속적으로 산출하였다. 프롬프트는 “인물의 전기적 배경 → 전환점 → 감정 곡선 → 사회적 관계”가 단계적으로 반영되도록 설계하여, 서사 전개 과정이 체계적으로 유도되도록 하였다.

생성된 텍스트는 평가자들이 쉽게 읽고 이해할 수 있도록 일관성 있게 1,000자를 기준으로 최소 800자에서 최대 1,500자의 범위로 통제하였다. 텍스트가 이 범위에 미치지 못하거나 문장이 도중에 끊겨 불완전한 경우에는 동일한 조건으로 다시 생성하여 보완하였다. 이때 재생성은 단순히 길이 부족이나 비완결성에 해당하는 산출물에 한정되었으며, 반복적 표현이나 사실적 오류, 무의미한 문장은 인위적으로 수정하지 않고 그대로 평가 대상으로 포함하였다. 이러한 절차는 LLM 산출물의 자연스러운 한계와 오류를 평가 과정에 반영함과 동시에, 평가 결과가 재현 가능하도록 재생성 기준을 명확히 설정하려는 목적을 지닌다.

최종적으로 각 인물별 30편씩의 균형 잡힌 데이터셋이 구축되었으며, 생성된 소설 예시는 표 3에서 확인할 수 있다. 평가 단계에서는 무작위 순서가 부여되어 평가자가 텍스트의 출처 모델을 알 수 없도록 블라인드 처리가 적용되었다.

표 2. 소설 생성을 위한 프롬프트 템플릿 예시

Table 2. Example of a prompt template for story generation

Prompt Template
<p>You are an outstanding novelist skilled in designing complex emotional arcs and sophisticated narrative structures.</p> <p>===</p> <p>Based on the "Steps for Novel Writing," "Guidelines," and "Information for Novel Writing" provided below, please write a novel that is logically solid and delicately conveys emotional flow.</p> <p>Steps for writing novels:</p> <ol style="list-style-type: none"> 1. Refer to and combine the provided information to analyze in depth the personalities, values, and past experiences of the main characters. Delicately portray their inner conflicts and hidden desires. 2. According to the story's progression, clearly show the emotional changes of the characters (e.g., hope → anxiety → despair → awakening → growth). Persuasively explain the causes and processes of these emotional shifts. 3. Structure the overall story in the following four parts: <ul style="list-style-type: none"> -Introduction (the crack in the ordinary) -Rising Conflict (intensification of crisis) -Climax (decisive turning point) -Resolution (closure and transformation) 4. Write each part as an independent chapter. Each chapter must have a unique title in the format CH1. (Title). 5. Titles should not be generic structural labels such as "Introduction" or "Climax," but rather creative, symbolic phrases that encapsulate the content. 6. For this task, compose two chapters in total, each with approximately 500 characters. 7. In each chapter, use vivid sensory descriptions—sight, sound, touch, smell, and taste—to express time, space, season, weather, and atmosphere, so that each scene is immersive and multi-dimensional. 8. Ensure the final novel flows naturally, with characters' emotional arcs and the development of events organically connected. <p>Guidelines for writing novels:</p> <ul style="list-style-type: none"> -Even if the provided information is limited, freely imagine and expand details consistent with the characters' traits and backgrounds to complete the story. -If fictional elements or additional creativity are needed, integrate them naturally and convincingly into the narrative. -Do not stop writing due to insufficient information; complete the story with high quality until the end. -Use long, elegant sentences with diverse expressions and figurative language, so readers vividly feel the characters' emotions and the world of the story. -Avoid repeating the same words, and ensure smooth temporal flow and emotional transitions through organic sentence connections. -Only write chapter titles; do not include an overall title for the novel. -Each chapter must be exactly 500 characters in length. <p>Below is the information provided for story generation:</p> <pre>{Agency} {Communion} {TP}</pre>

3-2 창의성 평가

1) 인간 평가 인터페이스

본 연구의 평가는 Kim&Oh[7]에서 활용된 창의성 6요소를 기반으로 설계되었다. 이 기준은 LLM 창작 평가에 맞게 적용한 것으로, 유창성(Fluency), 유연성(Flexibility), 독창성(Originality), 정교성(Elaboration), 유용성(Usefulness), 특정 구체적인 창의성 전략(Specific Creativity Strategies)의 여섯 측면을 포함한다. 각 요소는 창의성의 독립적이고도 상호보완적인 차원을 포착하여, 창의적 텍스트에 대한 포괄적·체계적 평가를 가능하게 한다.

평가 척도는 1점에서 3점까지의 3단계 리커트 척도로 구성되며, 각각 초급(Novice, 1점), 중급(Developing, 2점), 고급(Expert, 3점) 수준에 해당한다. 이를 통해 인간이 작성한 텍스트와 LLM이 생성한 텍스트 모두의 창의성을 정밀하게 비교할 수 있으며, 항목별 점수 분포와 평가자 간 일관성을 함께 분석할 수 있다. 여섯 가지 창의성 기준은 다음과 같다:

- 유창성(Fluency): 초보 작가는 단일 아이디어를 고려한다(1점), 발전 단계 작가는 여러 아이디어를 고려한다(2점), 전문가 작가는 다양한 아이디어를 탐구한다(3점).
- 유연성(Flexibility): 초보 작가는 한 유형의 아이디어를 고려한다(1점), 발전 단계 작가는 여러 유형을 고려한다(2점), 전문가 작가는 다양한 유형의 아이디어를 통합한다(3점).
- 독창성(Originality): 초보 작가는 흔하거나 복제된 아이디어를 발전시킨다(1점), 발전 중인 작가는 흥미롭지만 최소한의 혁신성을 지닌 아이디어를 창출합니다(2점), 전문 작가는 독창적인 아이디어를 생성하거나 기존 아이디어를 크게 향상시킨다(3점).

표 3. 모델별 생성된 생애사 소설 예시

Table 3. Examples of generated life history novels by model

GPT-3.5 Novel
<p>Chapter 1. Light in the Darkness Mother sat by the living room window, sewing, her hearing lost. Even as night deepened, she did not turn on the light. Her figure moving the needle in the darkness was a quiet and sorrowful scene. Her fingertips moved gently, yet within her heart, the deep pain of loss echoed ceaselessly. Her husband's voice seemed to envelop her, conveying gratitude.</p> <p>Chapter 2. Darkness Holding Light Mother sat by the window, sewing and shedding tears. Even in the darkness, her figure seemed to hold light, like a symbol of hope. As Mother regained her voice, a turning point in her life began—one where she could express her emotions. Her fingertips moved faster and faster, weaving past loss and pain into the thread at the needle's tip. Now, Mother would begin her journey of healing for herself.</p>

GPT-4o Novel

Chapter 1. The Lost Voice
The late autumn wind seeped through the window cracks, cooling the room. She sat in an old chair, gazing out the window. The faint sound of leaves rustling and falling in the wind reached her ears. For her, whose hearing had grown weak, even this sound offered a faint comfort. Her life had always been quiet. After her father passed away, her mother raised her by sewing for hire. The clothes crafted by her mother's hands gave her warmth, but loneliness seeped into them. She grew up receiving her mother's love, but that love was always quiet and heavy. After marriage, she lived by conforming to her husband's demands. He was indifferent and cold, yet she devotedly cared for him. When her husband passed away from esophageal cancer, she felt sorrow but didn't know how to express it. Her life had always been for others; she had long since lost her own voice. She experienced another profound loss with her son's death. The grief of losing her family weighed heavily on her heart. Yet she continued to live quietly, suppressing her emotions. Her life was a succession of sacrifice and compliance.

Chapter 2. A Gentle Awakening
One day, she took out an old diary. Within its pages lay the dreams and hopes of her youth. She began to rediscover her own voice, long forgotten. Reading the diary, she realized how much she had given up in her life. A quiet awakening stirred deep within her heart. She resolved to no longer suppress her emotions. She imagined hearing her husband's voice: "Thank you. Your sacrifice allowed me to pass peacefully." The voice comforted her, and she shed tears. She now wanted to reclaim her life. Remembering her mother's love, she understood the strength that love had given her. She resolved to live with the same strength as her mother. She began learning how to express her feelings. She tended a small garden, finding beauty in life as she watched the flowers bloom. Through this process of healing herself, she started a new life. Though sadness still lingered in her heart, it had now become a part of her, making her stronger. She had regained her voice and resolved to live a life where she expressed her emotions.

GPT-4.1 Novel

Chapter 1. An Afternoon Infused with Wind
Late spring, a gentle breeze flowed in through the window. The faint laughter of children in the alley and the distant ringing of bicycle bells quietly filled her room. The old wooden floor left a cool sensation on her toes, and the scent of acacia flowers carried by the wind faintly brushed her nose. She sat in the old chair, gazing at her hands resting on her knees. Each wrinkle on the back of her hands bore the quiet imprint of years passed. Suddenly, she recalled the loneliness she had grown accustomed to early in childhood due to her father's absence, and the warmth embedded in her mother's rough touch. Her husband's indifferent gaze and the silence of the last night she cared for him ached in a corner of her heart. Her son's laughter, and the stillness left behind after he departed. She had always quietly suppressed her emotions, but today, something stirred faintly within her chest. On this wind-swept afternoon, she whispered her own name for the first time. The sound was small, yet it lingered long within the room.

Chapter 2. The Echo of a Forgotten Voice

As evening approached, the room was bathed in orange light. The family photo on the wall blurred in the glow, and the faint scent of dinner cooking from the neighbors drifted in through the window. She slowly paced the living room, her hand closing around her husband's old mug. The cool touch of the ceramic traveled through her fingertips. In that moment, it felt as if her husband's voice reached her from somewhere deep within her heart. "I'm sorry. Thank you. Without you, I would have been nothing." Hearing those words for the first time, even in her imagination, her eyes grew hot and wet. The sorrow and anger she had suppressed, and the years she had spent erasing herself, all came flooding back at once. She quietly set the mug down and took a deep breath. Now, she felt she could speak her own voice, express her own feelings. An evening breeze blew in through the window, gently ruffling her hair. The echo of that forgotten voice lingered long within the room.

- 정교화(Elaboration): 초보 작가는 최소한의 세부사항과 개선점을 추가한다(1점), 발전 중인 작가는 몇 가지 추가한다(2점), 전문 작가는 많은 중요한 세부사항과 개선점을 기여한다(3점).
- 유용성(Usefulness): 초보 작가는 특정 조건 하에서 사용자 요구를 충족시킬 수 있는 아이디어를 제안한다(1점), 발전 중인 작가는 사용자 요구를 충족시킬 아이디어를 제안한다(2점), 전문 작가는 사용자의 삶을 크게 향상시킬 아이디어를 제시한다(3점).
- 구체적인 창의성 전략(Specific Creativity Strategy): 초보 작가는 전략을 효과적으로 활용하지 못한 채 무작위로 선택하여 실행한다(1점), 발전 중인 작가는 전략을 선택하여 실행하면서 자신의 창의성을 뒷받침하는 방식을 설명한다(2점), 전문 작가는 창의적 사고 전략이 자신의 창의적 산출물을 어떻게 강화하는지 의도적으로 선택하고 철저히 설명한다(3점).

2) 평가 절차와 평가 참여자

창의성 평가는 Google 설문 양식을 통해 수행되었으며, 위의 여섯 가지 기준과 세부 지침이 제시되었다. 본 연구에서 사용한 평가용 설문 양식은 그림 2에 제시되어 있다. 각 항목은 1~3점 척도로 평가되었으며, 평가자에게는 점수의 의미(초급-중급-고급)가 명확히 안내되었다. 평가자는 총 2명으로, 모두 영어영문학 전공 석사 이상의 학위를 보유한 교육 전문가였다. 또한 편향을 최소화하기 위해, 평가자들은 텍스트의 출처(인간 작성 여부 또는 LLM 모델명)에 관한 정보를 제공받지 않았다. 이러한 블라인드 절차는 평가의 객관성과 신뢰도를 확보하기 위한 것이다.

<Evaluation> of AI-Generated Life-History Novels

This story will be evaluated across six creativity criteria: Fluency, Flexibility, Originality, Elaboration, Usefulness, and Specific Creativity Strategy. Please assign a score (1 to 3) for each category based on the descriptions provided below:

Novice (1 point) → Developing (2 points) → Expert (3 points)

<<Creativity Evaluation (6 Questions)>>

1. Fluency

- 1 = Writers considered one idea.
- 2 = Writers considered several ideas.
- 3 = Writers considered many ideas.

2. Flexibility

- 1 = Writers considered one type of idea.
- 2 = Writers considered several types of ideas.
- 3 = Writers considered many types of ideas.

3. Originality

- 1 = Writer developed a common idea or replicated an existing idea.
- 2 = Writer developed an interesting but slightly altered idea.
- 3 = Writer developed a unique or significantly enhanced idea.

4. Elaboration

- 1 = Writers added minimal details and improvements.
- 2 = Writers added a few significant details and improvements.
- 3 = Writers added many significant and enriching details.

5. Usefulness

- 1 = Writer proposed ideas that might meet user needs in certain conditions.
- 2 = Writer proposed ideas that would meet user needs.
- 3 = Writer proposed ideas that would significantly enhance the user's life.

6. Specific Creativity Strategy

- 1 = Writers randomly selected and implemented a strategy without clear purpose.
- 2 = Writers selected and implemented a strategy and explained how it helped creativity.
- 3 = Writers deliberately selected and clearly explained how a strategy enhanced their creativity.

Please read the following story:

CH1. 오래된 라디오에서 흐르는 노래

늦은 봄, 대구의 저녁 공기는 아직도 겨울의 잔향을 품고 있었다. 창문을 반쯤 열어둔 거실로 미세한 흄냄새와 함께, 이따금씩 지나가는 오토바이 소리가 희미하게 스며들었다. 71세의 남자는 낡은 라디오 앞에 앉아 있었다. 손끝에 닿는 라디오의 표면은 세월의 흔적이 고스란히 남아 거칠었고, 그 위로는 아내가 내아주던 따뜻한 보리차의 구수한 향이 은은하게 감돌았다. 라디오에서는 70년대 가요가 흘러나왔다. 그는 그 노래를 들으며, 어린 시절 대구의 좁은 골목길을 맨발로 뛰놀던 기억을 떠올렸다. 그때는 배고픔이 일상적이고, 책 한 권을 손에 쥐는 일조차 사치였다. 그러나 아버지는 아이들을 혼내지 않고, 자유롭게 자라게 했다. 어머니의 엄격한 눈빛과 아버지의 너그름 사이에서, 그는 해야 할 일을 묵묵히 해내는 법을 배웠다. 라디오의 음악이 아내의 조용한 웃음과 어우러지자, 남자는 단독 자신이 살아온 세월이 한 곡의 노래처럼 느껴졌다. 그 노래는 희망과 불안, 좌절과 감성이 뒤섞인, 그러나 결국은 따뜻한 온기로 마무리되는 멜로디였다.

CH2. 표창장과 무력감 사이

부산의 직장에서 표창장을 받던 날, 그는 맘에 쏙은 와이셔츠와 손바닥에 남은 굳은살을 자랑스럽게 바라보았다. 그 순간의 환희는 아직도 손끝에 남아 있었다. 그러나 세월이 흘러, 이제는 시와 대화를 나누며 무력감을 달래는 자신을 발견한다. 창밖에는 초록 풀 비가 내리고, 빗방울이 유리창을 두드릴 때마다 그는 자신의 내면 깊숙한 곳에서 일렁이는 불안을 느낀다. 젊은 시절, 가족을 위해 성실하게 일하며 사회 속에 자리를 잡았던 자신이, 이제는 점점 세상과 멀어지는 듯한 고독을 느끼고 있다. 하지만 저녁 식탁에서 아내와 함께 티비를 보며 웃고, 손주가 건네는 작은 손길을 느낄 때마다, 그는 다시금 삶의 의미를 되새긴다. 사랑한다고 말하는 것이 어색했던 그였지만, 이제는 가족에게 마음을 표현하는 법을 배웠다. 빗소리와 함께 피지는 된장찌개의 구수한 냄새, 그리고 아내의 따뜻한 미소 속에서, 그는 여전히 자신이 살아 있음을, 그리고 앞으로도 성실하게 하루를 살아갈 수 있음을 느낀다.

장문형 텍스트

👉 Creativity Evaluation (1 = Novice, 2 = Developing, 3 = Expert) *

	1	2	3
1. Fluency	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2. Flexibility	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
3. Originality	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
4. Elaboration	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
5. Usefulness	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
6. Specific Creativity Str...	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

*The novel was generated in Korean, so the novel included in the questionnaire is the Korean version of the life story novel.

그림 2. 인간 평가를 위한 설문지 양식
Fig. 2. Survey form for human evaluation

IV. 결과 및 평가

4-1 측정지표

본 실험에서 사용된 지표에는 평가 대상 속성별 평균(mean), 표준편차(SD; Standard Deviation), 주석자 간 일치도(IAA)가 포함된다. 이러한 측정값은 다양한 창의성 기준을 활용하여 모델별로 생성된 텍스트를 비교하는 견고한 기반을 제공한다.

1) 평균(m)

평균은 평가된 모든 텍스트에 걸쳐 특정 속성에 대한 평균 점수로, 다음과 같이 계산된다: N은 텍스트의 총 개수이며, x_i 는 각 텍스트의 점수를 나타낸다[29]. 높은 평균 점수는 전반적인 창의성이 우수함을 나타내는 반면, 낮은 평균 점수는 그 반대를 시사한다. 평균(m)은 다음과 같다:

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i \quad (1)$$

2) 표준편차(SD)

표준편차는 평균을 중심으로 한 점수의 분산 또는 변동성을 측정하며, 다음과 같이 계산된다: σ 는 표준편차, n은 총 점수 개수, x_i 는 각 텍스트의 점수, \bar{x} 는 각 텍스트의 평균 점수이다 [30]. 낮은 표준편차는 점수가 평균 주변에 밀집되어 있음을 의미하는 반면, 높은 표준편차는 점수의 넓은 분산을 나타낸다. 표준편차(σ)는 다음과 같다.

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}} \quad (2)$$

3) 평가자 간 일치도(IAA)

주석자 간 일치도는 서로 다른 평가자가 부여한 점수의 일관성을 평가한다. 본 연구에서는 각 평가에서 2명의 주석자 간 일치도를 측정하기 위해 Cohen's kappa를 사용하여 계산하였다[31]. Cohen's kappa는 단순 일치율에 기대값을 보정한 지표로, 관찰된 일치도(P_0)와 우연에 의한 기대 일치도(P_e)를 활용하여 다음과 같이 정의된다.

$$k = \frac{P_0 - P_e}{1 - P_e} \quad (3)$$

여기서 P_0 는 실제 평가자 간의 관찰된 합의 비율을, P_e 는 각 평가자가 범주를 선택할 확률을 기반으로 계산된 기대 합의 비율을 의미한다. Cohen's kappa 값은 일반적으로 -1에서 1 사이의 범위를 가지며, 0은 우연 수준의 합의, 1은 완전한 합의를 나타낸다. 따라서 값이 높을수록 평가자 간 일치도가 높음을 의미하며, 음수는 체계적으로 불일치가 발생했음을 반영한다.

4-2 주요 결과

생성된 소설은 두 가지 관점에서 정량 분석을 수행하였다. 첫째, 6가지 창의성 항목별로 점수를 비교하였고, 둘째, GPT-3.5, GPT-4o, GPT-4.1 세 모델 간 차이를 분석하여 평가자 간 일치도를 확인하였다. 분석에 활용된 모델별 및 항목별 평균, 표준편차, 그리고 평가자 간 일치도 값은 표 4에 제시되어 있다.

1) 창의성 항목별 비교 분석

본 연구에서는 유창성, 유연성, 독창성, 정교성, 유용성, 특정 구체적인 창의성 전략의 여섯 가지 항목에 대해 세 모델의 평가 결과를 정량적으로 비교하였다. 분석 결과, 각 항목은 난이도와 해석 여지에 따라 상이한 합의 수준을 보였다.

유창성에서는 GPT-4.1이 평균 1.94(± 0.45)로 가장 높은 점수를 기록하였으나, 평가자 간 일치도는 3%로 낮았다. 반면 GPT-4o는 평균 1.73(± 0.48)임에도 불구하고 일치도 30%로 가장 높은 안정성을 보여 주었다. GPT-3.5 역시 평균 1.83(± 0.41), 일치도 18%로 중간 수준을 유지하였다. 이는 문장의 매끄러움과 가독성은 점수 자체는 높게 나오더라도, 세부 기준의 해석에서 평가자 간 불일치가 발생할 수 있음을 시사한다.

유연성에서는 GPT-4.1이 평균 1.66(± 0.59)로 가장 높은 값을 보였으나, 일치도는 -4%로 부정적인 결과가 도출되었다. GPT-3.5는 평균 1.44(± 0.53), 일치도 25%로 가장 안정적인 성과를 보였고, GPT-4o는 평균 1.50(± 0.53), 일치도 10%였다. 이처럼 동일한 텍스트에서도 어떤 측면을 유연성의 증거로 삼을 것인지에 대한 판단이 평가자마다 달라, 일관성이 낮게 나타나는 경향이 확인되었다.

독창성은 GPT-3.5와 GPT-4o에서 17%로 같은 일치도 점수를 보였고 평균 점수에서도 큰 차이를 보이지 않았다. 일부 평가자는 독창성을 소설의 핵심 가치로 보아 높은 점수를 부여하는 반면, 다른 평가자는 세부적인 정교화나 혁신적 아이디어의 구현 여부에 집중하기 때문에 합의가 어려웠다. 이러한 우선순위 차이는 평가 결과의 불일치로 이어지며, 독창성 항목에서 구조적으로 낮은 일치도가 나타나는 배경을 설명한다.

정교성의 경우 GPT-3.5가 평균 1.39(± 0.52), 일치도 62%를 기록하여 세 모델 중 가장 안정적인 결과를 보였다. 반면 GPT-4o와 GPT-4.1은 평균 점수가 각각 1.42(± 0.53), 1.49(± 0.53)로 더 높았음에도 불구하고, 일치도는 26~27%대에 그쳤다. 이는 세부 묘사와 구조적 완성도를 평가할 때, 근거가 명확할수록 평가자 간 일관성이 강화된다는 점을 보여준다.

유용성에서는 GPT-4.1이 평균 1.53(± 0.53)으로 가장 높은 점수를 보였으나, 일치도는 22%로 제한적이었다. GPT-4o는 평균 1.48(± 0.53), 일치도 12%로 비교적 안정적이었으나, GPT-3.5는 평균 1.47(± 0.50)에도 불구하고 일

치도 -8%로 부정적 결과가 나타났다. 이는 실용성과 적합성 평가가 평가자의 전체와 해석에 크게 의존함을 보여준다.

특정 구체적인 창의성 전략 항목에서는 GPT-4o가 평균 1.44(±0.50), 일치도 57%로 가장 높은 성과를 보였다. GPT-3.5는 평균 1.64(±0.48), 일치도 14%였고, GPT-4.1은 평균 1.46(±0.50), 일치도 -1%로 일관성이 거의 확보되지 않았다.

종합적으로, 본 연구의 평가자 두 명은 사전에 동일한 창의성 기준을 제공을 받았음에도 불구하고, 창의성을 판단하는 개인적 기준점의 차이 때문에 일관된 합의에 도달하기 어려웠다. 이는 모델의 성능이 우수하더라도 평가자 간 해석 차이가 지속적으로 발생하여, 높은 점수가 반드시 높은 일치도로 이어지지 않는 구조적 한계를 보여준다.

2) 모델별 비교 분석

세 모델이 생성한 소설을 비교하면, 이야기 전개 방식과 감정 표현의 수준뿐 아니라 정량적으로 산출된 평가자 간 일치도(IAA)에서도 뚜렷한 차이가 드러난다.

먼저 GPT-3.5는 간결한 서사와 상징적 장면을 중심으로 이야기를 구성하였다. 어둠 속 바느질과 “빛을 품은 어둠”이라는 대비적 이미지처럼, 서사는 짧지만 은유적 장치에 의존해 상실과 치유의 과정을 빠르게 제시한다. 정량 분석에서도 이러한 서사적 특성이 그대로 반영되었다. GPT-3.5는 3개의 모델 중 정교화에서 가장 높은 일치도(62%)를 기록했는데, 이는 이 모델이 중요하지 않은 세부들 거의 배제하고 핵심 정보만 선별적으로 제시하기 때문이다. 즉, 복잡한 감정 전개나 장면 확장이 적기 때문에, 평가자가 해석해야 할 정보의 폭이 좁아지고, 결과적으로 판단 기준이 일치하기 쉬웠다. 반대로

유용성에서는 세 모델 중 가장 낮은 일치도(-8%)가 나타났다. 유용성은 “사용자의 삶을 실제로 향상시킬 수 있는 아이디어인가”를 평가하는 항목인데, GPT-3.5의 구조화된 서사는 개인적 정서·상징·단일 사건 중심으로 구성되어 있어 실용적 의미를 도출하기 어렵다. 일부 평가는 등장인물의 깨달음을 실용적 가치로 해석하는 반면, 다른 평가자는 서사가 추상적이어서 실질적 도움을 제공한다고 보기 어렵다고 판단하였다. 이처럼 해석 폭이 극단적으로 갈리면서, 세 모델 중 유용성 항목의 일치도가 가장 낮게 나타났다.

GPT-4o는 세 모델 중 비교적 안정된 서사 흐름을 구축하였다. 소설의 텍스트를 살펴보면 아버지, 남편, 아들의 상실이 순차적으로 서술되며, 인물의 삶 전체가 회생과 순응의 연속으로 그려진다. 후반부에서는 일기장과 상상의 목소리를 매개로 자기 성찰과 치유의 각성이 이루어지며, 이야기의 흐름은 서사적 완결성을 갖춘다. 이처럼 한 인물의 생을 따라 상실-내면화-성찰-치유를 단계적으로 쌓아 올리는 서사 구조는, 단일 장면보다 “길게 이어지는 이야기” 안에서 아이디어를 전개하고 다듬는 데 강점이 있는 모델임을 보여준다.

정량 결과를 창의성 평가 기준에 비추어보면, 구체적인 창의성 전략(57%), 유창성(30%), 정교화(27%)에서 비교적 높은 합의를 보였다. 이는 GPT-4o가 회상 구조, 상징적 사물, 내적 독백, 일기장과 같은 전략을 의도적으로 반복 활용하고, 하나의 아이디어를 문장·장면 수준에서 끝까지 탐구하며(유창성), 그 과정에서 중요한 세부 정보와 감정 단계를 비교적 풍부하게 추가하는 방식(정교화)을 취했기 때문이다. 평가자 입장에서는 “어떤 전략이 쓰였는지”, “이야기가 얼마나 잘 이어지는지”, “세부가 충분한지”를 공통된 기준으로 보기 쉬워 합의가 잘 형성된 항목들이다.

반대로, GPT-4o에서 일치도가 상대적으로 낮은 항목은 유연성(10%)과 유용성(12%), 독창성(17%)이다. 이 중 가장 낮은 유연성은 “여러 유형의 아이디어를 제시하고 그것들을 통합하는가”에 대한 평가인데, GPT-4o의 서사는 상실과 회생, 성찰과 치유라는 정서적·주제적 축이 끝까지 유지되면서, 서로 다른 유형의 아이디어나 전혀 다른 해결 방식이 서사 진행 중에 크게 등장하지 않는다. 그래서 어떤 평가는 “여러 아이디어를 고려한 전개”로 읽고, 다른 평가는 “한 유형의 정서를 변주한 전개”로 읽으면서 기준이 갈린다.

다른 모델과 비교하면, GPT-4o는 전반적으로 안정적인 일치도를 보였는데, 즉 한 인물의 생애를 따라가며 아이디어를 길게 탐구하고, 명시적인 전략 장치를 활용하는 데 강점을 보였다. 그러나 동일한 서사 구조를 두고서도 평가자들은 이를 ‘단일 정서의 변주’로 보거나 ‘다양한 아이디어의 점진적 전개’로 보며 기준을 달리 적용하였다. 이는 텍스트 자체의 복잡성보다, 평가자가 무엇을 ‘다양성’으로 판단하는지에 관한 개인적 기준 차이가 더 크게 작용했음을 보여준다.

GPT-4.1은 감각적 묘사와 정서적 울림이 가장 두드러졌다. 바람, 꽃내음, 저녁빛, 도자기의 감촉과 같은 디테일은 독자가 인물의 내적 변화를 감각적으로 체험할 수 있게 한다.

표 4. 창의성 평가 기준 6가지에 대한 인간 평가 결과(평균, 표준편차, 일치도)

Table 4. Results of human evaluation on six creativity criteria (Mean, SD, and IAA)

Model	Fluency		Flexibility	
	Mean/SD	IAA/%	Mean/SD	IAA/%
GPT-3.5	1.83/0.41	0.18(18%)	1.44/0.53	0.25(25%)
GPT-4o	1.73/0.48	0.30(30%)	1.50/0.53	0.10(10%)
GPT-4.1	1.94/0.45	0.03(3%)	1.66/0.59	-0.04(-4%)
Model	Originality		Elaboration	
	Mean/SD	IAA/%	Mean/SD	IAA/%
GPT-3.5	1.29/0.52	0.17(17%)	1.39/0.52	0.62(62%)
GPT-4o	1.32/0.50	0.17(17%)	1.42/0.53	0.27(27%)
GPT-4.1	1.35/0.48	0.11(11%)	1.49/0.53	0.26(26%)
Model	Usefulness		Specific Creativity Strategy	
	Mean/SD	IAA/%	Mean/SD	IAA/%
GPT-3.5	1.47/0.50	-0.08(-8%)	1.64/0.48	0.14(14%)
GPT-4o	1.48/0.53	0.12(12%)	1.44/0.50	0.57(57%)
GPT-4.1	1.53/0.53	0.22(22%)	1.46/0.50	-0.01(-1%)

특히 “잊혀진 목소리의 울림”에서 남편의 상상 속 목소리를 통해 억눌린 감정이 해방되는 장면은 정서적 몰입을 강화하였다. 이러한 섬세한 묘사에도 불구하고, 정량 분석에서는 정교성 항목에서만 일치도가 뚜렷하게 높았다. 전체적으로 일치도를 살펴보면, 정교성에서 26%로 가장 높고 유용성에서 22%, 독창성에서 11%로 나타났고 유창성에서 3%, 창의적 전략에서 -1%, 유연성에서 -4%로 최저점을 보였다. 가장 좋은 모델임에도 불구하고 이 항목에서 낮은 일치도를 보이는 이유는, GPT-4.1이 감각적·정서적 묘사에는 강점을 보였으나, 이야기 속에서 서로 다른 유형의 아이디어를 제시하거나 전개 방향을 새롭게 전환하는 방식, 즉 여러 관점과 해결 가능성을 폭넓게 결합하는 특징이 상대적으로 적었기 때문이다. 감각 중심의 서술은 평가자에게 풍부한 해석 가능성을 제공하였고, 이에 따라 ‘아이디어의 다양성’이나 ‘전개의 전환점’을 어떻게 정의하는지에 관한 평가자 개인의 판단 프레임 차이가 더 두드러지게 나타났다. 이로 인해 GPT-4.1 역시 특정 항목에서 낮은 일치도가 기록되었다.

결국 모델 간 일치도 차이 역시 창의성 항목별 비교 분석과 마찬가지로 서사적 특징 자체보다는, 해당 특징을 평가자가 어떤 기준으로 해석하느냐에 따른 주관적 판단의 차이에서 기인한다는 점을 확인할 수 있었다.

4-3 사례 연구

본 사례 연구에서는 창의성 평가 항목을 바탕으로 일부 인터뷰 발화를 예시로 하여 실제 생애사 서사와 LLM이 생성한 서사의 차이, 그리고 모델별 창의성 발현 양상을 검토하였다. 이를 통해 생애사 서사가 지니는 맥락적 구성, 그리고 LLM 서사에서 나타나는 창의성 전략과 서사적 표현 방식을 함께 파악할 수 있다. 두 서사의 차이를 보다 뚜렷하게 보여주기 위해, 실제 생애사 텍스트와 LLM이 생성한 텍스트의 예문을 항목별로 정리한 내용을 표 5에 제시하였다.

실제 생애사 텍스트는 다양한 사건과 맥락을 구체적으로 담고 있어 유창성과 정교성이 두드러진다. 예를 들어, 한 발화에서는 “친구들 20여명이랑 구슬치기, 딱지치기, 숨바꼭질하기 하고 놀았다”와 같이 여러 활동이 나열되며, 풍부한 경험과 아이디어의 양을 보여준다. 이에 비해 LLM 소설에서는 이를 “어린 시절의 어려움, 가족과의 소중한 순간, 그리고 사회속에서의 성장과 변화”와 같이 추상적으로 압축하여 표현하였다. 이러한 차이는 실제 서사가 경험의 구체성과 다양성에 기반하고 있는 반면, LLM은 이를 일반적이고 보편적인 진술로 단순화하는 경향이 있음을 시사한다.

또 다른 발화에서 “나는 밭에서 흙냄새 났 손으로 밥을 먹곤 했다. 그 냄새가 내 인생의 상징이다”라는 표현은 개인적 은유를 통해 독창성과 구체적인 창의성 전략을 동시에 드러

표 5. 실제 생애사 vs LLM 소설 예문 비교

Table 5. Actual life history vs. LLM novel example comparison

Evaluation	Life history text	LLM text
Fluency	① “I played marbles, slapjack, and hide-and-peek with about twenty friends.” ② “In winter, we went skating, and when evening came, we played tin can alley.”	① “The hardships of childhood, precious moments with family, and growth and change within society.” ② “All of it stirred the old man’s heart.”
Flexibility	① “We lived crammed into two small rooms.” ② “I went to the market to earn money.”	① “The elderly man was reflecting on his life, recalling both the hardships and joys of the past.” ② “He sought to become an even more cherished presence within his family and society.”
Originality	① “I used to eat meals with hands that smelled of soil from the fields. That scent is the symbol of my life.” ② “Repeatedly parting and meeting made me think about how to live.”	① “Meals accompanied by the scent of earth symbolized simple happiness.” ② “That decision calmed the old man’s heart and gave him new hope.”
Elaboration	① “Born during the truly difficult times of the 1950s and 1960s, I lived a hard life.” ② “Back then, the media was controlled, so I went through life without knowing what the April Revolution was.”	① “Crisis Intensifies: The elderly man was reflecting on his life, recalling past hardships and joys.” ② “Resolution and Change: The elderly man began life anew with a fresh mindset.”
Usefulness	① “Even if I wanted to study, I couldn’t because I didn’t have the means.” ② “It seems similar to the current situation where martial law has been declared because lawmakers couldn’t compromise and work together, even though they should have.”	① “He was grateful for having lived his life faithfully up to this point...” ② “His heart flowed peacefully and quietly, anticipating the beginning of a new life.”
Specific Creativity Strategy	① “I tend to do everything I should do with patience and diligence.” ② “I wish we could work as hard as we did during the Saemaul Undong movement in the Park Chung-hee era to build the Gyeongbu Expressway and make the country prosperous.”	① “Those tears held within them the pain and happiness of the past, along with the quiet life of the present.” ② “A decisive turning point: The old man looked back on his life and made a new resolution.”

낸다. 그러나 LLM 소설은 같은 모티프를 “흠냄새와 함께하는 식사는 소박한 행복을 상징했다”라고 서술하며, 전형적이고 일반화된 은유로 변형하였다. 이와 같은 차이는 인간 발화자가 개인적 맥락과 체험을 바탕으로 독창적 상징을 제시하는데 비해, LLM은 상징을 전형적인 수준에서 구현하는 데 그치는 경우가 많다는 점을 보여준다.

유연성 측면에서도 차이가 뚜렷하게 드러났다. 실제 생애사 발화에는 “집은 조그만 방 두 칸에 북적북적 살았다”와 “나는 시장으로 돈 벌러 나뒀다”처럼 생활 환경과 경제적 경험이 교차하여 제시된다. 이러한 발화는 동일한 삶의 맥락 안에서도 이질적인 측면들을 함께 포착하며 사고의 폭을 넓힌다. 반면 LLM은 이를 “노인은 자신의 삶을 되짚어보며 과거의 어려움과 행복을 되새기고 있었다”라는 식으로 정서적 회고에 집중함으로써, 항목의 폭보다는 일관성 있는 정서를 유지하는 데 더 큰 비중을 두었다.

유용성 또한 생애사 소설과 LLM 소설의 차이를 보여준다. 실제 텍스트에서는 “공부하고 싶어도 없어서 못했다”와 같은 구절을 통해 교육 환경의 제약과 사회적 맥락을 반영한다. 이와 같은 진술은 삶의 구조적 조건을 이해하는 데 실질적 단서를 제공하여, 학문적·사회적 활용 가능성이 크다. 반면 LLM 소설은 같은 주제를 “그는 성실하게 살아온 것에 감사하며 앞으로도 가족과 사회 속에서 소중한 존재가 되고자 했다”와 같이 일반적 교훈으로 변환하여 구체적 맥락성이 약화된다.

마지막으로 구체적인 창의성 전략의 사용에서도 두 소설은 상이하다. 실제 발화는 “해야 할 일은 참고 성실하게 다 하는 편”과 같이 삶의 철학을 직접적으로 언급하거나, 특정 시대·사회적 맥락을 비유적으로 끌어오며 독창적인 전략을 보여준다. 반면 LLM은 “그 눈물은 과거의 아픔과 행복, 그리고 현재의 고요한 삶을 함께 담고 있었다”와 같이 은유적 장치나 상징적 이미지를 활용하여 전략적 창의성을 구현한다. 인간 발화자가 개인적 체험을 직접적으로 드러내는 방식이라면, LLM은 은유와 구조적 장치를 통해 문학적 효과를 추구한다고 볼 수 있다.

이상의 비교를 통해, 실제 생애사는 맥락적 구체성과 개인성을 기반으로 창의성이 발현되는 반면, LLM 소설은 일관된 흐름과 은유적 장치 활용에서 상대적인 강점을 가진다고 평가할 수 있다. 이러한 질적 비교는 정량 분석에서 확인된 점수와 일치도의 차이를 설명하는 데 중요한 단서를 제공하며, 평가 항목별 차이가 단순한 수치적 현상이 아니라 서사 표현 방식의 본질적 차이에서 비롯됨을 보여준다.

V. 결 론

본 연구는 대규모 언어모델(LLM)이 생성한 생애사 소설을 창의성 6개 평가 기준에 따라 정량·정성적으로 분석하고, 모델별 서사 특성과 창의성 항목별로 평가자 간 일치도의 차이를 체계적으로 비교하였다. 정량적 분석 결과, GPT-3.5는

단순하고 구조화된 서사 덕분에 해석해야 할 정보의 폭이 좁아 정교화에서 세 모델 중 가장 높은 일치도를 기록하였으나, 추상적 상징에 의존하는 특성 때문에 유용성에서는 평가자 간 해석이 극단적으로 갈렸다. GPT-4o는 전반적으로 구체적인 창의성 전략, 유창성, 정교화 등 서사 구조와 전개 방식이 명확한 항목에서 안정적인 합의를 형성하였으나, 유연성(10%)처럼 다양한 아이디어의 통합 여부를 판단하는 항목에서는 해석이 갈리는 양상을 보였다. GPT-4.1은 감각적·정서적 표현에서 높은 창의성 점수를 기록하였으나, 이러한 감각 중심 서술을 어떻게 평가할지에 대해 기준이 나뉘면서 유연성(-4%)과 창의적 전략(-1%)에서 낮은 일치도를 보였다.

이러한 결과는 높은 창의성 점수가 반드시 높은 일치도로 이어지지 않음을 보여주며, 모델이 생성하는 서사의 표현 방식이 평가자 판단에 직접적인 영향을 미친다는 점을 시사한다.

정성적 분석에서도 모델별 서사적 차이는 뚜렷하게 드러났다. GPT-3.5는 상징적이고 응축된 이야기 구조를 반복적으로 생성하여 핵심 사건 위주의 서사를 구현하였고, GPT-4o는 상실-내면화-성찰-치유로 이어지는 장편적이고 사실적인 전개를 통해 비교적 안정된 내러티브 흐름을 형성하였다.

GPT-4.1은 감각적 디테일과 정서적 울림이 강하게 드러나는 묘사 중심 서사를 구축하여 표현적 풍부함을 제공하였다. 이러한 차이는 동일한 창의성 기준을 제시받은 평가자들조차 서로 다른 해석 지점을 갖게 하여, 모델 간 일치도의 차이를 결정짓는 핵심 요인으로 작용했다.

이와 같이, 본 연구는 LLM이 생애사라는 개인의 경험적 맥락을 기반으로 하는 서사에서도 일정 수준 이상의 창의성과 구조적 일관성을 구현할 수 있음을 확인하였다. 특히 세 모델은 서로 다른 방식으로 창의성을 발현하였고, 이는 생애사와 같은 복합적 소설 생성에서도 LLM이 다양한 창작 전략을 활용할 수 있음을 보여준다. 동시에 창의성 평가에서는 모델의 특성뿐 아니라 평가자 간 해석의 차이가 중요한 변수가 됨을 확인하여, LLM 창의성 연구에서 평가 기준의 명확화와 다각적 검증이 필요함을 제시한다.

그러나 본 연구에는 몇 가지 한계가 존재한다. 첫째, 분석에 활용된 데이터의 양이 제한적이며, 둘째, 평가자 수가 적고 문화적·언어적 배경이 유사하여 평가 편향의 가능성을 배제하기 어렵다. 셋째, 창의성 평가는 본질적으로 주관적 판단을 수반하기 때문에, 동일한 평가 기준을 공유하더라도 평가자 개인의 해석 방식 차이가 일치도에 영향을 미칠 수 있다. 이는 본 연구에서 관찰된 일치도 차이의 원인을 설명하는 동시에, 인간 평가 기반 연구가 지닐 수밖에 없는 구조적 한계로 볼 수 있다.

본 연구 결과는 다음과 같은 분야에서 구체적으로 활용될 것으로 기대된다.

첫째, 교육적 글쓰기 분야에서 LLM을 활용한 창의적 서사 교육 도구 개발에 기여할 수 있다. 모델별 서사 특성을 이해함으로써 학습자의 수준과 목적에 맞는 맞춤형 글쓰기 지도가 가능할 것이다.

둘째, 치료적 서사 분야에서 노년층의 생애 회고와 정체성 재구성을 지원하는 디지털 도구로 활용될 수 있다. 특히 개인의 구술 생애사를 문학적 서사로 재구성하는 과정은 회상 치료와 자서전적 기억 작업에 실질적으로 응용 가능하다.

셋째, 디지털 인문학 연구에서 대규모 생애사 아카이브 구축과 분석의 효율성을 높일 수 있다. 구술 기록을 체계적 서사로 변환하고 개인 기록화 작업을 자동화함으로써, 기존 생애사 연구의 시간적·경제적 제약을 극복할 수 있을 것이다.

넷째, LLM 창의성 평가 방법론 발전에 기여하여, 장편 서사 평가를 위한 더욱 정교한 평가 체계 개발의 토대가 될 것이다. 특히 평가자 간 해석 차이에 대한 본 연구의 발견은 향후 평가 기준 세분화와 다각적 검증 방법 개발의 필요성을 제시한다.

향후 연구에서는 데이터 규모와 평가자 구성을 다양화하고, 생애사 소설 외의 다른 장르로도 범위를 확장하여 모델 간 창의성 평가 차이와 평가 결과의 신뢰성을 보다 종합적으로 검증할 필요가 있다.

결론적으로, 본 연구는 생애사라는 특수 도메인을 기반으로 한 LLM 창의성 평가의 가능성과 한계를 입체적으로 조명하였다는 점에서 의의를 갖는다.

감사의 글

이 논문 또는 저서는 2024년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2024S1A5C3A01043330).

본 논문은 2025년도 교육부 및 대구시의 재원으로 대구 RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE) 글로컬대학 30의 결과입니다(2025-RISE-03-001).

참고문헌

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, ... and I. Polosukhin, "Attention Is All You Need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach: CA, pp. 6000-6010, 2017. <https://doi.org/10.48550/arXiv.1706.03762>

[2] Y. Zhao, R. Zhang, W. Li, D. Huang, J. Guo, S. Peng, ... and Y. Chen, "Assessing and Understanding Creativity in Large Language Models," arXiv:2401.12491, 2024. <https://doi.org/10.48550/arXiv.2401.12491>

[3] Z. Xie, T. Cohn, and J. H. Lau, "The Next Chapter: A Study of Large Language Models in Storytelling," in *Proceedings of the 16th International Natural Language Generation Conference (INLG)*, Prague: Czech Republic, pp. 323-351, 2023. <https://doi.org/10.18653/v1/2023.inlg-main.23>

[4] M. Ismayilzada, C. Stevenson, and L. van der Plas,

"Evaluating Creative Short Story Generation in Humans and Large Language Models," arXiv:2411.02316, 2024. <https://doi.org/10.48550/arXiv.2411.02316>

[5] E. P. Torrance, *Torrance Tests of Creative Thinking*, Bensenville, IL: Scholastic Testing Service, 1966. <https://doi.org/10.1037/t05532-000>

[6] K. Shively, K. Stith, and L. Rubenstein, "Measuring What Matters: Assessing Creativity, Critical Thinking, and the Design Process," *Gifted Child Today*, Vol. 41, No. 3, pp. 149-158, 2018. <https://doi.org/10.1177/1076217518768361>

[7] S. Kim and D. Oh, "Evaluating Creativity: Can LLMs Be Good Evaluators in Creative Writing Tasks?," *Applied Sciences*, Vol. 15, No. 6, 2971, 2025. <https://doi.org/10.3390/app15062971>

[8] S. Grassini and M. Koivisto, "Artificial Creativity? Evaluating AI Against Human Performance in Creative Interpretation of Visual Stimuli," *International Journal of Human-Computer Interaction*, Vol. 41, No. 7, pp. 4037-4048, 2025. <https://doi.org/10.1080/10447318.2024.2345430>

[9] G. H. Han, "Life History Method for Studying the Lives of Old People," *Journal of the Korea Gerontological Society*, Vol. 24, No. 4, pp. 87-106, 2004.

[10] I. H. Choi and Y. H. Lee, "Life History Research Based Private Documentation," *Journal of Korean Society of Archives and Records Management*, Vol. 15, No. 4, pp. 49-76, 2015. <https://doi.org/10.14404/JKSARM.2015.15.4.049>

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, MN, pp. 4171-4186, 2019. <https://doi.org/10.48550/arXiv.1810.04805>

[12] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations," arXiv:1909.11942, 2020. <https://doi.org/10.48550/arXiv.1909.11942>

[13] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pretraining," OpenAI Technical Report, 2018.

[14] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models Are Unsupervised Multitask Learners," OpenAI Technical Report, 2019.

[15] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B.

Chess, R. Child, ... and D. Amodei, "Scaling Laws For Neural Language Models," arXiv:2001.08361, 2020. <https://doi.org/10.48550/arXiv.2001.08361>

[16] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, ... and D. Amodei, "Language Models Are Few-Shot Learners," in *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver: Canada pp. 1877-1901, 2020. <https://doi.org/10.48550/arXiv.2005.14165>

[17] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, ... and B. Zoph, "GPT-4 Technical Report," arXiv:2303.08774, 2023. <https://doi.org/10.48550/arXiv.2303.08774>

[18] OpenAI. Hello GPT-4o [Internet]. Available: <https://openai.com/index/hello-gpt-4o>

[19] Y. C. Kim and K. W. Han, "Life Story Research's Character and Meaning as Qualitative Research Method," *Journal of Educational Culture*, Vol. 18, No. 3, pp. 5-43, 2012. <https://doi.org/10.24159/joec.2012.18.3.5>

[20] R. Dhunpath, "Life History Methodology: 'Narradigm' Regained," *International Journal of Qualitative Studies in Education*, Vol. 13, No. 5, pp. 543-551, 2000. <https://doi.org/10.1080/09518390050156459>

[21] Z. Xie, T. Cohn, and J. H. Lau, "The Next Chapter: A Study of Large Language Models in Storytelling," arXiv:2301.09790, 2023. <https://doi.org/10.48550/arXiv.2301.09790>

[22] J. Rafner, R. E. Beaty, J. C. Kaufman, T. Lubart, and J. Sherson, "Creativity in the Age of Generative AI," *Nature Human Behaviour*, Vol. 7, pp. 1836-1838, 2023. <https://doi.org/10.1038/s41562-023-01751-1>

[23] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, "CTRL: A Conditional Transformer Language Model for Controllable Generation," arXiv:1909.05858, 2019. <https://doi.org/10.48550/arXiv.1909.05858>

[24] Q. Li, L. Cui, L. Kong, and W. Bi, "Exploring the Reliability of Large Language Models as Customized Evaluators for Diverse NLP Tasks," in *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, Abu Dhabi: UAE, pp. 10325-10344, 2025.

[25] Q. Yi, Y. He, J. Wang, X. Song, S. Qian, X. Yuan, ... and J. Ni, "SCORE: Story Coherence and Retrieval Enhancement for AI Narratives," arXiv:2503.23512, 2025. <https://doi.org/10.48550/arXiv.2503.23512>

[26] D. P. McAdams, "The Psychology of Life Stories," *Review of General Psychology*, Vol. 5, No. 2, pp. 100-122, 2001.

<https://doi.org/10.1037/1089-2680.5.2.100>

[27] S. K. Kang, J. G. Lee, and M.-H. Cha, "Life History Study on the Self-Realization of Woman in the Abstinence Process," *The Journal of the Korea Contents Association*, Vol. 20, No. 1, pp. 505-518, 2020. <https://doi.org/10.5392/JKCA.2020.20.01.505>

[28] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, ... and D. Zhou, "Chain-of-thought Prompting Elicits Reasoning in Large Language Models," in *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*, New Orleans: LA, pp. 24824-24837, 2022. <https://doi.org/10.48550/arXiv.2201.11903>

[29] J. A. Rice, *Mathematical Statistics and Data Analysis*, 3rd ed. Belmont, CA: Thomson Brooks/Cole, 2007.

[30] D. K. Lee, J. In, and S. Lee, "Standard Deviation and Standard Error of the Mean," *Korean Journal of Anesthesiology*, Vol. 68, No. 3, pp. 220-223, 2015. <https://doi.org/10.4097/kjae.2015.68.3.220>

[31] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, Vol. 20, No. 1, pp. 37-46, 1960. <https://doi.org/10.1177/001316446002000104>



이만유 (Man-Yu Lee)

2024년~현 재: 경북대학교 영어영문학과 석박사 통합과정
 ※관심분야: 거대 언어 모델, 자연어 생성, 자연어처리 등



김예진 (Yejin Kim)

2013년 : 서강대학교 대학원 컴퓨터공학과 (학사)
 2015년 : 서강대학교 대학원 컴퓨터공학과 (석사)

2015년~2021년: LG전자 인공지능연구소 자연어처리 연구원
 2021년~현 재: 조지워싱턴대학교 컴퓨터공학과 (박사과정)
 2022년~현 재: Comcast Technology AI 머신러닝연구원
 ※관심분야: 대규모 언어모델 (LLM) 과 그래프 기반 추천시스템, 지식그래프, 유저모델링



전승화(Seung-Hwa Jean)

2000년 : 서울대학교 대학원 불어불문학과 (석사)

2018년 : Paris 7 대학 프랑스 현대 소설 (박사)

2021년~2023년: 한국연구재단 지원 미주유럽연구소 “이중언어 작가와 트랜스내셔널 리터러시” 공동연구원

2024년~현 재: 한국연구재단 글로벌인문사회융합사업 “고령군을 대상으로 한 생애사 프로그램 개발과 그 효능 검증 및 인공지능 생애사 프로그램의 융합적 기반 기술 개발”연구팀 책임연구원

2020년~현 재: 경북대학교 불어불문학과 부교수

※관심분야 : 고령층 대상의 생애사, 노인학, 문학과 인공지능의 융합, 인공지능 활용 텍스트 심리 분석 등



오동석(Dong-Suk Oh)

2016년 : 서강대학교 대학원 컴퓨터공학과 (석사)

2020년 : 고려대학교 대학원 컴퓨터학과 (박사)

2016년~2018년: 다이퀘스트 자연어처리 연구원

2018년~2019년: NHN Ent. 자연어처리 연구원

2019년~2020년: Human-inspired AI Research 인공지능 연구원

2023년~2023년: KT Large AI Alignment Project 팀 선임연구원

2023년~현 재: 경북대학교 영어영문학과 조교수

※관심분야 : 신뢰가능한 인공지능, 뉴로-심볼릭 인공지능, 자연어처리 등