



< 용 어 설 명 >

용어 및 약어	정의
sLLM	거대 언어 모델(LLM)중 상대적으로 크기가 작은 모델
환각현상(hallucination)	거짓 정보를 사실처럼 생성하는 생성형 모델의 오류중 하나
인스트럭션 튜닝 (instruction tuning)	미세조정(Fine tuning)의 방법 중 하나, 작업 및 목표를 input data로써 모델 학습에 포함시키는 방법
하이퍼파라미터 (hyper parameter)	AI에서 하이퍼파라미터란 학습 모델 훈련, 예측에 사용되는 외부 구성 변수. 수동으로 설정이 가능함
NDCG	Normalized Discounted Cumulative Gain - 모델이 예측한 순위를 반영한 추천시스템의 성능지표
schema.org	웹 페이지의 구조화된 데이터를 마크업하기 위한 표준 온톨로지로서 검색 엔진이 데이터를 더 잘 이해하고 활용할 수 있게 도와준다.
JSON-LD	Json을 사용하여 링크드 데이터를 인코딩 하는 방식
저작도구1	창작자를 위한 텍스트/음성 기반 저작도구. 창작자가 집필한 텍스트를 더 풍부하고 매끄럽게 집필할 수 있도록 도와준다. 목차 자동 생성과 편집 도구, 에피소드별 집필할 때 생성형AI로 원고 교열을 지원해주는 도구 등이 내장되어 있다.
저작도구2	텍스트 원고를 아티클(콘텐츠, 챕터) 단위로 나누어 DB로 제공하면 이용자가 원하는 아티클을 골라 엮어 새로운 책을 만들 수 있다. 주로 강사가 여러 교재에서 원하는 챕터를 골라 맞춤 교재로 엮을 때 사용한다. 컴복스가 2009년부터 챕터 PDF를 바탕으로 맞춤교재를 제작하는 서비스를 운영하고 있다. 문학작품의 경우 시, 단편소설이 하나의 단위가 될 수 있고, 인문사회과학 도서의 경우 챕터 단위가 아티클이 될 수 있다.
저작도구3	저작도구1에서 집필한 원고를 장르, 인물, 플롯 등 다양한 콘텐츠 요소를 중심으로 변경/전환하여 2차 창작을 할 수 있는 저작도구다. 창작자의 상상력을 증진하고, 다양하고 새로운 이야기를 생성할 수 있다.
아티클(article)	커뮤니케이션북스에서 정의한 용어. 도서의 상품화 가능한 가장 작은 단위. 대표적으로 ‘챕터’가 있으며, 문학 작품의 경우 시 한 편, 에세이 한 편이 아티클이 될 수 있다.
생성문학 (Generative literature)	생성형AI 기술을 활용하여 창작된 문학. 공동창작, 단독 창작 등 여러 방법이 있으며 아이디어, 플롯, 캐릭터, 글쓰기 등 집필과정에 생성형AI를 다양하게 활용한 결과물이다.
엔티티(entity)	데이터 설계에서 독립적인 단위로 속성을 가지는 가짐, 본 과제에서는 콘텐츠, 도서, 아티클이 가장 중요한 엔티티이다.
모티프(motif)	이야기에서 반복되는 주제나 상징, 개념을 나타내는데, 주로 인물과 사건의 관계임.
아카이빙	일반적인 DB 저장보다는 장기보존을 목적으로 하고 법적인 효력을 가질 수 있음.
스토리 아크(story arc)	이야기에서 감정과 긴장감의 변화를 그래프로 나타내면 클라이맥스에서 최고점을 가지는 곡선(arc)이 된다고 붙은 이름. 최근에는 여러 에피소드나 챕터가 연결된 큰 그림으로 이야기 전개를 가리키는 용어가 되었다. narrative arc라고도 함.
net4j 그래프 DB	그래프 데이터베이스 관리 시스템으로, 데이터와 관계를 노드와 엣지로 표현. Cypher 쿼리 언어를 사용하며, 실시간 처리와 복잡한 관계 분석에 강점.
Elasticsearch	실시간 분산형 검색 및 분석 엔진으로, 대량의 데이터를 빠르게 검색하고 분석 가능.

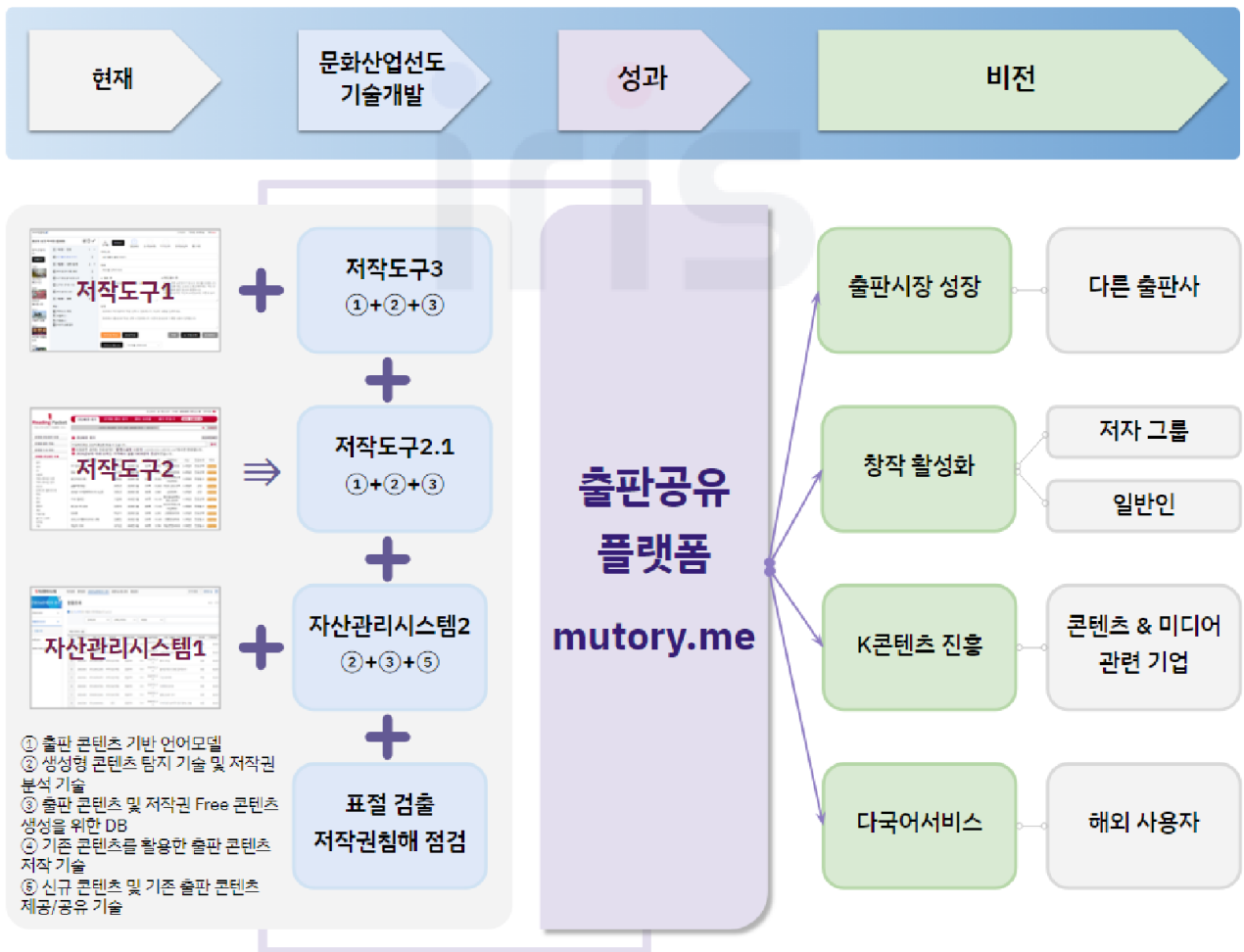
목 차		
본문1	1. 연구개발의 필요성	1
	1-1. 연구개발의 개요	1
	1-2. 연구개발 대상의 국내외 현황	2
	1-3. 연구개발의 중요성(필요성)	2
	1-4. 선행 연구 내용 및 결과	3
	2. 연구개발의 목표 및 내용	5
	2-1. 연구개발 최종목표	5
	2-2. 단계별 연구개발 목표 및 내용	7
	가-1. 1단계	7
	가-2. 1단계 - 개발목표	7
	가-3. 1단계 - 개발 내용	8
	나-1. 2단계	19
	나-2. 2단계 - 개발목표	19
	나-3. 2단계 - 개발내용	19
	2-3. 연구개발의 창의성·혁신성 등	23
	2-4. 연구개발 성능지표 및 평가방법	23
	가-1. 결과물의 성능지표	23
	가-2. 평가방법 및 평가환경	24
	나. 성과물 목표	24
	3. 연구개발의 추진전략·방법 및 추진일정	26
	3-1. 연구개발 추진전략·방법	26
	3-2. 연구개발 추진체계	27
	3-3. 연구개발 추진일정	28
	4. 마일스톤 체계 및 수행계획	29
	5. 연구개발성과의 활용방안 및 기대효과	32
	5-1. 연구개발성과의 활용방안	32
	5-2. 기대효과	33
	5-3. 사업화 전략 및 계획	33

1. 연구개발의 필요성

1-1. 연구개발의 개요

- ‘유일성’은 출판과 콘텐츠, 저작권 가치의 본질임.
- 출판을 뿌리로 생성형AI 기술을 접목하여 ‘나의 특별한 이야기’에서 시작하는 출판 콘텐츠 공유 서비스 플랫폼 ‘mutory.me’를 구현하여 지식 다양성과 접근성을 실현하고자 함.
 - 왜 나의 이야기인가?
 - 모든 콘텐츠는 스토리의 확장물이다.
 - 모든 스토리는 ‘인간 이야기’의 확장물이다.
 - ‘나의 이야기’는 유일한 경험이다.
 - ‘경험의 유일성’은 ‘창작물의 원천’이다.
- 주관사는 현재 ‘자서전 자동출판서비스(저작권구1)’(2023), ‘맞춤형교재 리딩패킷서비스(저작권구2)’(2009)를 운영중이며, ‘제품 자산관리시스템’을 자체 보유하고 있음.
- 본 연구과제를 통해 ‘생성형AI 기반 저작권구3’과 ‘저자/계약 자산관리시스템’을 추가 개발하고, 저작권 침해 방지와 공유 기능을 탑재한 ‘출판 콘텐츠 공유 서비스 플랫폼’을 론칭하려 함.

mutory.me
My Unique Story
나의 특별한 이야기



- 주관사는 ‘지식 다양성’을 사업 가치로 삼아 ‘세상에 읽을 수 없는 책이 없게’ 하는 사명을 실현하는 출판사. 출판의 새로운 역사를 쓰는 혁신력을 핵심역량으로 주제출판, 다품종 시장지배, 미디어 익스텐션과 롱테일 머천다이징 전략으로 다각도의 R&D를 통해 출판업의 성장 비전을 모색해 왔음.

1-2. 연구개발 대상의 국내외 현황

기술현황 및 시장현황

구분	국내	국외
기술현황 및 전망	• 생성형 AI 기반 스토리 창작에 대한 관심이 높아지고 있는 가운데, 문학시장에서도 ‘생성문학’(Generative literature)에 대한 창작과 연구, 기술 개발이 시작되고 있는 단계. • 어린이용 도서(그림책, 동화책 등)에서도 ChatGPT 등을 이용한 스토리 구성과 이미지 생성 등이 활용. • 웹툰과 웹소설 등 스토리 기반 콘텐츠를 쉽게 창작하고 유통할 수 있게 돕는 인터랙티브 콘텐츠 플랫폼 시장 활성화.(zeta)	• 생성형 AI를 통한 스토리 창작의 기술 개발과 세부 프로세스 연구가 상당히 진척되어 있음. • 서사학, 사회학, 심리학, 역사학 등 기존 인문사회과학의 성과를 AI 기반 스토리 생성을 위한 내러티브의 계획과 구조, 규칙, 변용에 활용. • Novel. AI 등 생성형 AI 기반 소설 창작 플랫폼이 상용화 • 인지 심리학과 서사학의 스토리 서스펜스에 대한 두 가지 이론적 기초에 기초한 새로운 반복 프롬프트 기반 계획 방법을 제안.
시장현황 및 전망	• 출판시장 2022년 매출액은 4조 5966억 • 국내 AI시장 24년 3조 662억원 규모로 예상 • 2023년부터 연평균 14.9% 성장 • 2027년까지 4조 4,636억원 규모	• 생성형 AI: 2022년 107억 9천만 달러 • 텍스트생성기: 2022년 3억 9,200만 달러 • 2032년까지 약 1,180억 6천만 달러 예상 • 2030년까지 연평균 17.3%의 성장률 예상

국내외 지적재산권 및 표준화

- 현재 AI 관련 특허 등록이 제약되는 추세에 따라 AI 창작 관련 다수 출원이 등록되지 않은 상태임.

구분	국내	국외
특허 등	• 저작권 관리 및 유통화 시스템, 발행 서버(저작권위원회) • 등록 저작권 정보 삽입 시스템 및 방법(저작권위원회) • 인공지능 기반의 소설 창작 도구 제공 시스템(출원중) • 2차창작 도구 관련(출원 3건) • 저작물 이력 관리 관련(출원 1건)	• (미국)AI-based content generation for gaming applications • (미국)Applied artificial intelligence technology for interactive story editing to support natural language generation (Narrative Science)(미국) • (미국) Applied artificial intelligence technology for runtime computation of story outlines to support natural language generation (NLG) 등
표준화	• 전자책 epub 2.0 • 저작권위원회 표절 관련 가이드라인	• schema.org를 통해서 웹문서로 음원, 영상, 소설 등을 공개 공유할 수 있는 파일 포맷 정의

1-3. 연구개발의 중요성(필요성)

생성형AI로 출판업 성장 저해 요인 타파 → 신규 성장동력 마련

- LLM 기반 언어 생성 기술은 출판업에 새로운 기회를 주었으나, 국내 출판사들의 영세한 규모(출판사2.5만개, 종사자 18.5만명, <2022한국출판연감>)와 기술력 부족으로 출판사 주체 사업적 활용은 거의 없음.
- 일부 IT업체가 전자책 콘텐츠나 온라인 콘텐츠 유통 차원에서 접근하고 있으나 출판업과 저작권에 대한 이해 부족으로 지속가능한 사업 모델을 제시하지 못하고 출판업계의 참여를 이끌어내지 못하고 있음.
- 일반 출판사는 종이책 중심 출판(전자책 출간 32.8% 불과) 도서 출간(인건비)과 제작(인쇄, 보관)에 비용이

많이 들어 예상 판매량이 높은 도서를 제한적으로 선택, **베스트셀러 중심의 출판 전략**으로 출간 종수(규모)가 제한되고 출판업 성장의 한계, **특히 디지털콘텐츠 비즈니스 전환의 한계**로 작용함.

- 현재 **출판사의 저작권 관리**는 ‘종이 계약서’와 ‘배타적 출판권’을 벗어나지 못해 **성장의 핵심 저해 요소**가 됨. 활용 가능한 저작권을 세부 정의하고 디지털화하여 온라인에서 자유롭게 활용될 수 있는 형태로 가공, 관리하는 시스템 필요.
- 예비 창작자의 작업물을 효율적으로 상품화하고 매출 기회로 연결할 시스템 필요.

1-4. 선행 연구 내용 및 결과

기관	연구 및 기술★	성과 요약	과제 관련성
커뮤니케이션 선복스 (주관)	commbooksrps.com 맞춤형 교재 리딩패킷 서비스	(2009년~현재) 도서 3845종, 챗터 27,750개 PDF 바탕으로 대학강사 대상 맞춤교재 1600종 제작 서비스. 15년 동안 사업 유지, 저자 사용료 정산.	본문 분할 사용과 2차 저작에 대한 저작물 이용권 확보와 운영 경험. 신사업 론칭, 운영 관리 노하우.
	자서전.net/main (2022, 중소기업기술정보 진흥원)	(2022~현재)성찰학습 기반 자서전 출판 자동화 서비스 AI기술 적용 개발. (특허출원10-2024-0064232) 저자의 인터뷰 답변에 기반한 자서전 자동 생성 시스템 및 그 방법	생성형AI 기반 출판 자동화 실현 자서전 교정교열 학습용 데이터 구축(특정 장르에 적합한 스타일) 기획 편집 자동화 프로세스 개발 국가 R&D 과제 주관 연구개발 경험
	인공지능 학습용 데이터 구축사업 (한국지능정보사회 진흥원)	(2022, 주관, 우수과제 선정)감성 및 발화스타일 동시 고려 음성합성 데이터, (참여) 다국어 통번역 데이터 (2021, 주관) 문학작품 낭송음성데이터	양질의 학습용 데이터 구축 경험 보유 출판도서 데이터별 100~500종씩 저작권자와 직접 계약 사용권 확보. 주관사로서 컨소시엄과 사업 관리 역량
	출판자산관리 시스템 (2022~현재, 자체 개발)	도서 6.2천종, 파생상품 5만 종(전자책, 큰글자책, 오디오북, 반값교재, 리딩패킷, 아티클 등) 관리용 자산관리 시스템 개발. 저자 4천명 관리, 계약 시스템 개발, DB 구축 중.	35년 업력 7개 브랜드 기반 다양한 장르 출판 콘텐츠와 신규 파생상품 사업화 경험 다양한 유형의 출판권, 2차적 사용, 2차 저작물 활용 생산 사업화 경험, 계약과 수익배분 시스템 설계, 개발, 운영 경험
	큰글자책 판매대행 사업 (2013~현재)	2013년부터 큰글자책 사업 시작, 2021년부터 타출판사 도서 판매 대행. 자사 4천종, 54개 출판사 1.2천종 판매. 무재고 주문생산 방식(2015 벤처기업 인증)으로 판매 종수 확대 중.	35년 업력과 출판계 네트워크, 기술력 바탕 사업 확대 중. “무재고 주문생산”으로 종이책 판매 종수 무한 증가 OSMU 실현
	자서전.net	① 구술 출판시스템 ② 자동 질문생성, 글 자동 교정 기능	1차 저작물 생성, LLM 기반 교정기능
바이칼AI (참여)	바른 형태소분석기	정확도 99.6%, 미등록단어 추출 문장분리, 띄어쓰기, 붙여쓰기 등 자동교정, 3,300이상 고객 사용	① 글의 띄어쓰기 등 오류 교정 기능 ② 표절시스템의 분석을 위해서 문장 단위, 의미단위 분할에 사용
	바른 개체명분석기	정확도 91.25%, 15개 개체명 인식 기능 보유	콘텐츠의 구성요소인 인물, 사건, 배경은 분석에 활용
	AI 달고나 (dalgona.ai)	① 학습 데이터 라벨링 도구 ② 정제, 가공 단계별 데이터 맞춤 편집 ③ 데이터 통계 및 권한 관리 기능	초기 학습 데이터에 대한 가공, 검수 과정에서 활용, 엔티티, 속성 등의 편집기로 활용가능
	메타 데이터	(2023.07~2023.12) KT 스카이라이프	메타 큐레이션 구축 사업화 경험
오투오	메타 데이터	(2023.07~2023.12) KT 스카이라이프	메타 큐레이션 구축 사업화 경험

기관	연구 및 기술★	성과 요약	과제 관련성
(참여)	큐레이션 구축	메타데이터 큐레이션 시스템 구축 완료 데이터 기반 추천 (통계, 메타), 알고리즘 추천 (AI), 편성자의 감성/전문성을 추가한 하이브리드 추천 적용	출판 데이터 DB 기반 메타데이터를 추출하는 작업과 연관 활용 가능 키워드의 카테고리별 분류 작업은 콘텐츠 구성 요소를 범주 별로 분리하는 과업과 매우 유사함
고려대학교 산학협력단 (참여)	대형 언어 모델 (KULLM)	기 개발된 sLLM 중 Upstage의 SOLAR 10.7B 모델을 기본 base 모델로 instruction-tuning하여 답변 정확도가 향상된 Chatbot형 모델 개발 데이터셋 (번역된 GPT4 ALL)으로 약 15만 건의 {instruction:output} 포맷의 데이터셋을 활용하여 학습	‘구름’ 모델 학습 과정은 sLLM 학습과 매우 큰 연관성을 가지고 있음 모델 학습에 필요한 학습 파라미터 선정, 최적화, 학습 데이터 전처리 등과 같은 과업에 있어서 매우 효율적인 진행 가능
	생성 AI 결과물의 Hallucination 기술 연구	신뢰할 수 있는 생성 AI 기술 확보를 위해 결과물에 대한 사실적 일관성과 진실성을 검증, 탐지하는 기술 연구, 환각 현상 완화 기술 개발(iitp 3개년)	환각 탐지와 완화에 검색 증강 생성을 사용하는 딥러닝 기술을 활용하여 출판 콘텐츠 특화 언어모델 개발에 기여할 수 있음

자산관리시스템1 - 자산관리시스템 (컴북스 내부 DB 구축 현황)

- 2024년 자체 개발 완료한 출판 자산 관리 시스템. 현재 54개 타출판사 자산 1.2천 종을 포함 자산 관리 중이며, 아티클 추가를 위해 1.5버전 개발 중.
- 제품 자산관리시스템에 이어 저자자산과 계약관리를 위한 자산관리시스템2 설계 개발 시작.

ID	제목	저자	출판사	상태
20023006	19711140222094 부파리들(가을)	권성자	다사	출판
20023005	19711140222097 부파리들(가을)	권성자	다사	출판
20023004	19711140222093 부파리들(가을)	권성자	다사	출판
20023003	19711140222096 부파리들(가을)	권성자	다사	출판
20023002	19711140222095 부파리들(가을)	권성자	다사	출판
20023001	19711140222098 부파리들(가을)	권성자	다사	출판
20023000	19711140222099 부파리들(가을)	권성자	다사	출판
20022999	19711140222100 부파리들(가을)	권성자	다사	출판

자산관리시스템 asset.commbooks.com

제목	저자	출판사	상태
19711140222094 부파리들(가을)	권성자	다사	출판
19711140222097 부파리들(가을)	권성자	다사	출판
19711140222093 부파리들(가을)	권성자	다사	출판
19711140222096 부파리들(가을)	권성자	다사	출판
19711140222095 부파리들(가을)	권성자	다사	출판
19711140222098 부파리들(가을)	권성자	다사	출판
19711140222099 부파리들(가을)	권성자	다사	출판
19711140222100 부파리들(가을)	권성자	다사	출판

자산관리시스템 제품 등록 화면

저작도구1 - 자서전.net

- 커뮤니케이션북스(주관), 바이칼시(참여) 공동 개발, 사용자가 직접 “자서전”을 창작하는 서비스

제목	저자	상태
19711140222094 부파리들(가을)	권성자	출판
19711140222097 부파리들(가을)	권성자	출판
19711140222093 부파리들(가을)	권성자	출판
19711140222096 부파리들(가을)	권성자	출판
19711140222095 부파리들(가을)	권성자	출판
19711140222098 부파리들(가을)	권성자	출판
19711140222099 부파리들(가을)	권성자	출판
19711140222100 부파리들(가을)	권성자	출판

제목	저자	상태
19711140222094 부파리들(가을)	권성자	출판
19711140222097 부파리들(가을)	권성자	출판
19711140222093 부파리들(가을)	권성자	출판
19711140222096 부파리들(가을)	권성자	출판
19711140222095 부파리들(가을)	권성자	출판
19711140222098 부파리들(가을)	권성자	출판
19711140222099 부파리들(가을)	권성자	출판
19711140222100 부파리들(가을)	권성자	출판

저작도구 2 - 리딩패킷 서비스 commbooksrps.com

- 2009년 개발 서비스 중인 맞춤형 교재 제작 서비스. 도서 3845종, 챕터 27,750개 PDF 바탕으로 대학강사 대상 맞춤교재 1600종 제작 서비스 중.



맞춤 교재 리딩패킷 서비스 commbooksrps.com



리딩패킷 만들기 기능 화면

2. 연구개발의 목표 및 내용

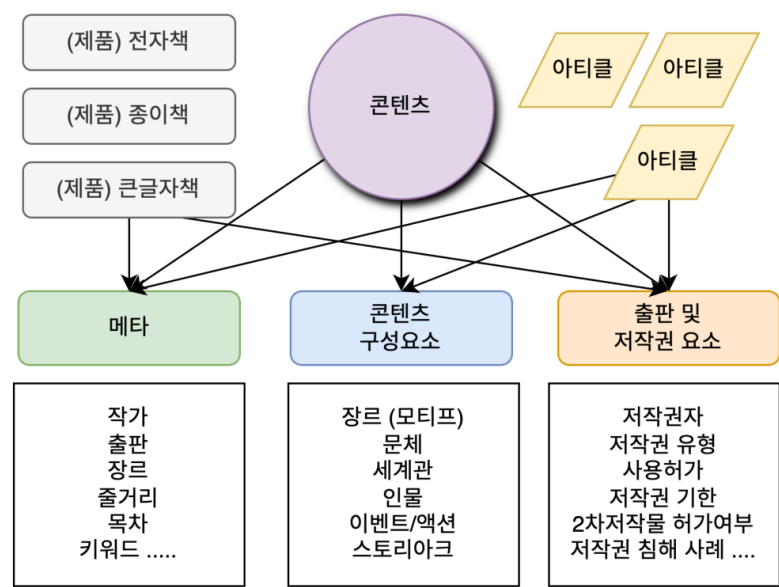
2-1. 연구개발 최종목표

- 출판 콘텐츠에 생성형AI 기술을 접목하여 '나의 특별한 이야기'에서 시작하는 출판 콘텐츠 플랫폼 'mutory.me'를 구현하여 서비스 론칭.
- mutory.me는 제품 자산관리시스템1, 저자/계약 자산관리시스템2와 3개의 저작도구(창작, 리믹스, 생성), 표절과 저작권 침해해 방지하는 기능을 탑재한 출판 콘텐츠 공유 플랫폼임.
- 5개 과제는 아래 그림과 같이 플랫폼 개발 과정에 적용됨.



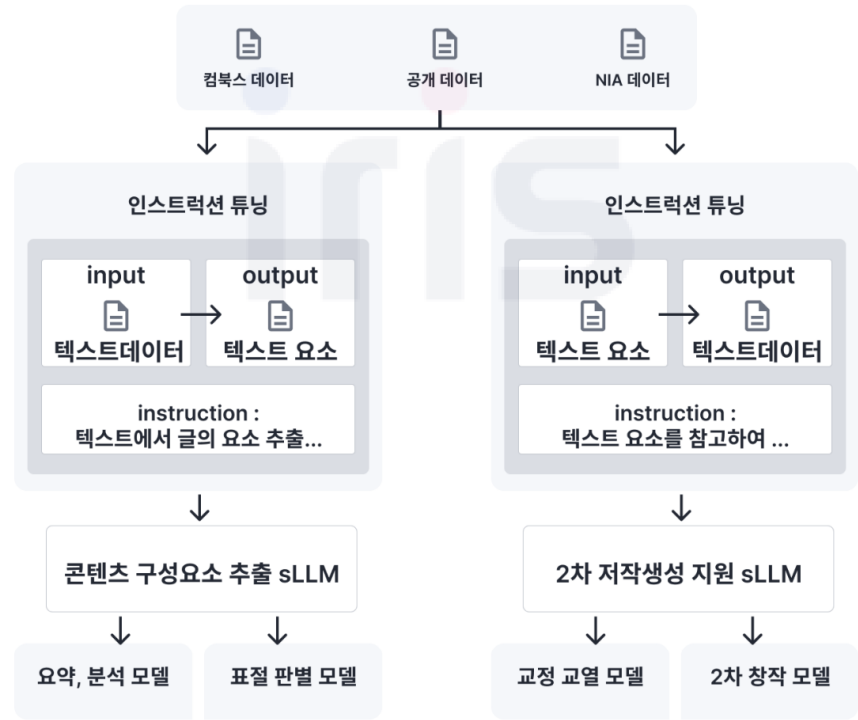
장르별 메타, 콘텐츠 구성 요소, 엔티티 구조도

- 당 과제에서는 엔티티(콘텐츠, 제품, 아티클) 단위를 아래 그림과 같이 정의함.
- 메타와 콘텐츠 구성 요소 및 출판 저작권 요소 관계도도 아래와 같음.



생성형 모델(LLM)과 하위 임무(Sub Task) 모델

- 당 과제에서 개발하는 LLM의 개념과 임무 개념도는 아래와 같음.



연구개발 세부 목표와 산출물

No	과제	담당	세부 내용	산출물 및 목표
1	출판 콘텐츠 기반 언어모델	컴복스	- 생성형 “★저작도구3”을 위한 장르 메타정보 정의	컴) 장르 메타정보 정의서 1식 고) 출판콘텐츠추출모델 1개 고) 2차저작 생성지원 모델1개 고) 비SCI급 논문 2건, SCI급 논문1건
		고려대	- 출판콘텐츠 분석용 sLLM 개발	
2	생성형 콘텐츠	컴복스	- 콘텐츠 구성요소 정의	컴) 콘텐츠 구성요소 정의서 1식

No	과제	담당	세부 내용	산출물 및 목표
	탐지 기술 및 저작권 분석 기술	오투오	- 콘텐츠 요소 분석 기술 개발 - 콘텐츠 표절 여부 탐지 기술 개발	오) 콘텐츠 요소 분석 AI 서비스 모듈 오) 콘텐츠 표절 여부 AI 탐지 모듈
3	출판 콘텐츠 및 저작권 Free 콘텐츠 생성을 위한 DB 구축 기술	컴복스	- 장르 선정과 학습데이터 구축 - 원천 자료 데이터화	컴) 학습용 파인튜닝 데이터 3천개 컴) 원천 자료 데이터화 3만개
		바이칼	- 장르별 엔티티 & 메타정보활용 DB 구축 - 장르별 엔티티 & 메타정보 관리도구 개발 - 저작권 침해요소별 분석 및 아카이빙 - 침해요소 케이스 DB 공유 기술	컴) 세부 속성정보 DB 1식 바) 엔티티&메타정보활용 DB 1식 바) 엔티티 DB관리 관리도구1식 바) 저작권침해요소 분석결과 보고서 1식
4	기존 콘텐츠를 활용한 출판 콘텐츠 저작 기술	컴복스	- 콘텐츠 구성요소 메타정보 생성 기술 분석	바) 요소 추천 서비스모듈 1식 바) 교정교열 AI 서비스모듈 1식 바) 2차 저작 AI 서비스모듈 1식
		바이칼	- 메타,요소 활용 저작도구 개발 - 요소추천 알고리즘 개발 - 메타기반 교정 교열 기술 개발 - 관계형 동작정의 기반 저작 기술 개발	
5	신규 콘텐츠 및 기존 출판 콘텐츠 제공/공유 기술	컴복스	- 저작권 침해 데이터 정의, 공유서비스 모델을 위한 DB 구축(저자/계약/수익배분)	컴)저작권 침해요소 아카이빙 컴) 콘텐츠공유플랫폼 기획안 1식
		바이칼	- 공유서비스 아키텍처 설계 - 콘텐츠 공유 플랫폼 개발 - 저작권 침해요소 케이스 분석 DB 개발	바) 콘텐츠 공유 플랫폼 1식(사용성 평가) 바) 저작권 침해요소 케이스 분석DB 1식

2-2. 단계별 연구개발 목표 및 내용

가-1. 1단계

설계	▶	사업 방향을 반영하여 장르 선정과 구성요소, 메타정보 정의 등
모델	▶	콘텐츠 추출 및 생성을 위한 모델 설계 및 구현 요소분석 및 추출, 교정/교열, 추천, 2차 창작 모델 개발
데이터	▶	콘텐츠 생성을 지원하기 위한 출판 콘텐츠 DB화
서비스	▶	출판 공유 플랫폼 (초기버전)

가-2. 1단계 - 개발목표

유형	기관	개발목표
주관	커뮤니케이션북스	③ 콘텐츠 생성 지원 DB 구축 / 장르 선정 및 초기 데이터 확보 ① 출판언어모델 / 장르 메타정보 정의 ② 탐지분석기술 / 콘텐츠 구성요소 정의, 검증 ④ 저작기술 / 콘텐츠 구성요소 메타정보 생성 기술 분석 ⑤ 콘텐츠 제공&공유기술 / 출판콘텐츠 상품특성 분석, 저작권 침해 데이터 정의

유형	기관	개발목표
공동1	고려대학교 산학협력단	① 출판언어모델 / 콘텐츠 분석용 sLLM개발 - 출판콘텐츠 구성요소 분석 개발
공동2	오투오	① 텍스트 데이터 메타데이터 추출 - 각 텍스트 데이터 당 메타데이터 10건 내외 추출 ② 텍스트 데이터 요약/분석 모델 개발 - 추출적 요약 기술 개발, 추상적 요약 기술 개발, 통합 요약 시스템 개발
공동3	바이칼에이아이	③ 콘텐츠 생성 지원 DB 구축 / 원천자료 엔티티 및 메타정보 활용 가능 DB 구축 ④ 저작기술 / 메타정보, 엔티티, 요소간 저작도구 개발, 요소추천 알고리즘개발, ⑤ 콘텐츠 제공&공유기술 / 공유서비스 아키텍처 설계, 콘텐츠 공유 서비스 개발

가-3. 1단계 - 개발 내용

가. 주관연구개발기관 - 커뮤니케이션북스

장르 선정과 학습데이터 구축 [③ 콘텐츠 생성 지원 DB 구축]

- 장르 선정 및 초기 데이터 확보
 - 장르(genre)**, 모티프(motif), 문체(style), 세계관(fictional universe), 인물(character), 이벤트/액션(event/action), 스토리 아크(story arc). “2차 창작을 위한 요소 정의”에 자세한 내용 참조. 각 분야 전문가 자문받아 수행.
- 장르별 학습 데이터 선정 및 저작권 확보
 - 7개 브랜드 저자 3,700명, 문학작품 2,500여 종 보유. 국가(언어), 시대별 다양성 고려하여 선정.
 - 컴북스는 리딩패킷 서비스(2009~)와 인공지능학습용데이터 구축사업(2021~2022, NIA) 참여로 이미 다수 타이틀에 대한 사용권을 확보했음.

소설 후보 작품과 저작권 확보			*저작권을 확인한 타이틀은 텍스트 선정 작업장에 등록합니다.			
ISBN	도서명	저자명	지자/각색/지작권자	지자사망년도	국가별	지만지발행일
9788964069806	보래사나미	오프만, 에른스트 I. A.	권역순(북문역)	1822	독일	2011-11-07
9788966803194	조판본강경애작품집	강경애	김경수(서강대)	1943	한국	2012-04-17
9788966805631	성암투안의유혹	귀스타브 플로베르	김계선	1880	프랑스	2012-10-25
9788966803415	아쿠타가와 류노스케 단편집	아쿠타가와 류노스케	김명주	1927	일본	2012-05-25
9791130461625	마키노 신이치 단편집	마키노 신이치	김명주	1939	일본	2015-02-25
9788966802883	조판본백신에 단편집	백신에	김문주	1939	한국	2012-03-29
9791128855085	사기꾼방랑여인쿠라세의인생기	한스 폰 그리멜스하우젠	김미란	1676	독일	2020-08-15
9791128834196	레스코프중단편집	니콜라이 레스코프	김민수	1895	러시아	2018-02-27
9791128857010	캠프케 중단편집	갈리나 캠프케	김민수	2019	러시아	2021-06-28
9791130427737	홍당무	쉴 르나르	구(저작재산권자 3년김희)	1910	프랑스	2016-10-25
9788966806386	박물지	장화	김영식(중)	300	중국	2013-01-21
9788964067758	열이전	조비	김장환	0	중국	2011-07-22
9788966803637	서경잡기	유홍	김장환	23	중국	2012-11-30

콘텐츠 구성요소 메타 정보 생성 기술 분석 [④ 저작기술]

- 위에서 정의한 콘텐츠 구성요소를 참여기업인 오투오의 ‘메타데이터 추출 기술’을 활용하여 콘텐츠에서 자동 추출할 수 있는지 테스트함.
- 목표 데이터 3만 건의 1~5%를 샘플링하여 자동 추출 결과 검토, 적절성 판단.

출판 콘텐츠 특성 분석 [⑤ 제공 및 공유 기술]

콘텐츠 단위 저작권에 따른 특성 분석

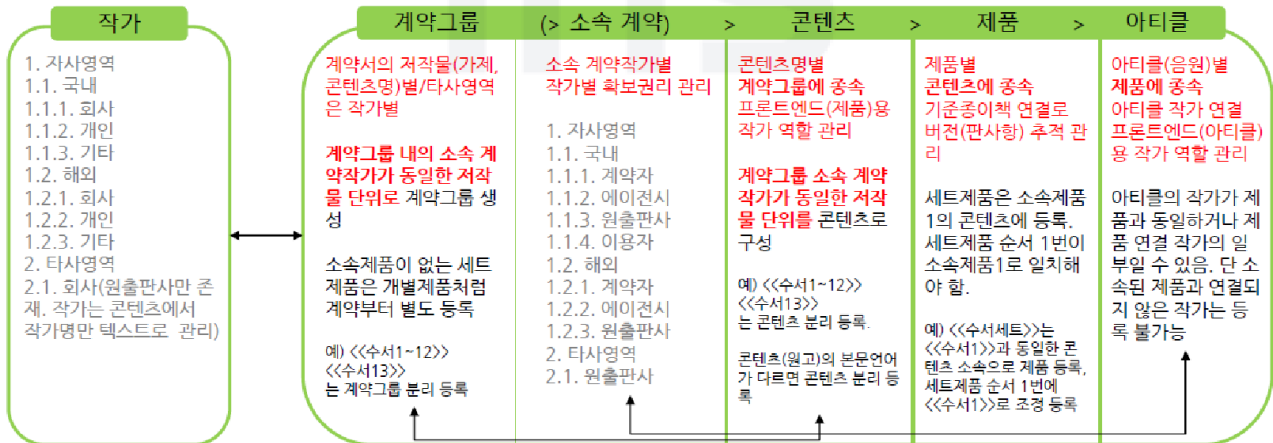
- 콘텐츠를 텍스트와 이미지(사진, 삽화, 표/그래프), 기타 요소로 분류하여 저작권 개별 관리 여부 판단.
 - 텍스트는 기본적으로 저작권자의 저작물로 관리. 단, 다른 저작물에서 부분 인용이 있을 경우 별도의 저작물로 표시하여 관리 필요함.
 - 사진, 삽화, 표/그래프는 창작, 인용과 구입 등 여러 경로로 삽입되었을 가능성 있음.
- 저작권자의 창작물이 아닌 경우 아티클(혹은 구성요소) 단위로 개별 등록하여 저작권 관리.

제품 단위 콘텐츠 특성 분석

- 상품으로서의 출판 콘텐츠 특성을 분석하고, 사업자의 권한 확보 여부에 따라 접근성과 활용을 제어.
- 콘텐츠와 제품 단위, 다양한 파생상품을 고려하여 저작권과 연계하여 다양한 상품의 자유로운 생산과 관리가 가능하도록 설계해야 함.
- 상품 대상은 기본 4가지, 사업 영역 확장에 따라 추가됨. ①종이책, ②큰글자책, ③전자책, ④아티클

상품 포맷	콘텐츠 데이터	유통채널	메타데이터 특징	판매 시 특징
종이책	PDF, 메타데이터	일반서점	제작정보 있음	무재고 주문생산, 익일 출고
큰글자책	PDF, 메타데이터	일반서점	제작정보 있음	무재고 주문생산, 익일 출고
전자책	Epub 2.0, 메타데이터	전자책서점	쪽수 없음	전문 서점 등록, DRM 관리
아티클	PDF, Epub 2.0, 메타데이터	플랫폼	도서정보 + 아티클 정보	플랫폼 등록, 판매

- 저작권을 바탕으로 콘텐츠 > 제품 > 아티클 단위로 설계 운영 예정임(하단 그림 참조).



저작권 침해 데이터 정의 [⑤ 제공 및 공유 기술]

- 저작권은 저작권자가 다른 이의 저작권을 침해하지 않았음을 보장함. 출판사는 저작권자와의 계약으로 사용권을 획득. 출판사는 크리에이티브 커먼즈 라이선스(creative commons license; SA,BY,NC,ND)와 저작권 계약 조항(22조)에 맞춰 권한 지정하여 활용.

저작권 침해요소 케이스 분석을 위한 저작권의 분해

출판저작물	저작권자	이용자	권한	기간	재산권	인격권적
문학, 드라마,	개인 혹은 단체	무료이용	라이선스	한시적 이용권	복제권	공표권

출판저작물	저작권자	이용자	권한	기간	재산권	인격권적
영화, 웹툰, 웹소설 등	창작자 양도받은자	정당한 권리 획득	양도 공정이용(범위 제한)		배포권 전시권 공중송신권 2차저작물작성권	성명표시권 동일성유지권

- 위 표에 따라서 확보한 라이선스의 유형에 따라서 허용되는 사용자 권한에 의거하여 이용자의 권한에 대해서 문제가 없는지 확인하여, 위반의 유형을 케이스로 정리함.
- **생성형AI를 활용한 창작물에 저작권을 인정할까?** 창작자의 사상과 의도가 포함될 경우 저작권 발생. 창작 과정(기획, 편집)에서 사상과 의도가 들어가도록 설계.

2차 창작물의 저작권 관리와 침해 방지 방안

- 저작자의 2차 창작 과정에서 **의도하지 않은 저작권 침해**를 일으킬 수 있으므로, 이를 방지하기 위해 ‘저작도구’ 개발 단계부터 침해를 방지하도록 설계.
 - (저작도구1) 1차 창작 과정에서 여러 번의 원고 수정과 교정/교열 과정을 거치게 하여 새로운 창작물이 되도록 설계함. 집필 과정에 생성형AI를 활용, 이용자가 주도적으로 전체 구성, 편집, 수정하므로 저작물로서 요건 충족, 새로운 저작권 발생함.
 - (저작도구2) 2차 창작 가운데 리믹스북 창작용 DB는 정식 계약된 콘텐츠만으로 구축함. 서비스에서 편집한 콘텐츠는 각각의 저작권자에게 저작권료를 지급하는 새로운 공동창작물로 출판, 저작권 사용료도 자동으로 계산 배분됨.
 - (저작도구3) 생성형AI 모델도 공개 DB나 저작권을 확보한 학습 데이터만 활용하여 제작함. 저작도구 활용 창작 과정에서 이용자가 수정/편집하여 저작권이 발생함.
 - 모든 창작물은 저작물로 등록된 후에 공유, 활용이 가능함. 저작물로 등록되지 않은 콘텐츠는 창작에 활용할 수 없음. 저작물로 등록하는 순간 저작권자가 명확해지며 권한 책임도 명확해지고, 침해 여부의 판단도 명확해짐.
 - (표절 방지) 저작도구를 활용하여 창작된 콘텐츠를 자체 표절 검출 도구를 활용하여 크로스체크하고 유사도를 확인할 수 있음. 차후 분쟁 발생 시, 유사도를 침해 여부 판단과 배상금 책정에 활용할 수 있음.
- 이에 따라 mutory.me에서 사용되는 모든 콘텐츠는 정식으로 저작권 확보, 창작된 콘텐츠가 됨.

공유 서비스 모델을 위한 DB 구축 [⑤ 제공 및 공유 기술]

- 위에 정의한 콘텐츠 이용 범위(권한)에 따라 이용자가 자유롭게 활용할 수 있도록 DB와 플랫폼 구축
 - 콘텐츠 단위(콘텐츠 > 도서(제품) > 아티클) 정의하고
 - 각 단위별로 저작물 이용 범위와 조건 설정
 - 저작도구에서 조건에 따라 사용되며 2차 창작물로 재탄생

Ⅲ. 공동연구개발기관 - 고려대 산학협력단

sLLM 설계

- 콘텐츠 구성요소 추출 모델: 텍스트 데이터 내에서 장르, 모티프, 문체, 세계관, 인물, 이벤트, 스토리 아크를 추출하기 위한 모델임.

- 2차 저작 생성 지원 모델: 입력된 콘텐츠 구성 요소를 활용하여 새로운 텍스트를 생성.

콘텐츠 구성요소 추출 모델:

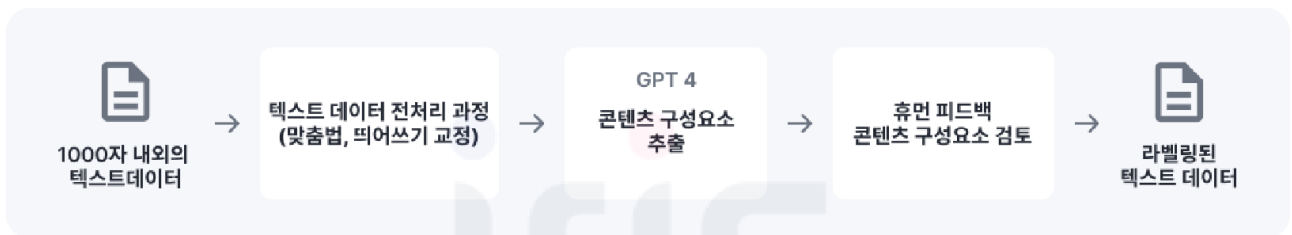
- 콘텐츠 구성 요소를 추출하기 위한 인스트럭션 데이터셋을 설계하고, 이를 바탕으로 콘텐츠 구성요소 추출 모델을 만드는 것이 주된 목표임
- 콘텐츠 구성요소 추출모델은 대량의 한국어 데이터를 사전 학습된 sLLM을 GPT 등 고성능 텍스트 생성 모델로 제작된 라벨링된 콘텐츠 요소 추출 모델의 학습용 데이터셋으로 instruction tuning하여 제작함
- 해당 과정은 전처리 과정 ▶ 데이터 라벨링 ▶ 모델 미세조정 순서로 이뤄짐.

2차 저작 생성 지원 모델:

- 주어진 콘텐츠 요소를 기반으로 창작물을 생성하는 모델 제작이 목표임.
- 창작물 생성 모델은 sLLM을 기반으로 만들어지며 첫번째 모델 학습 완료 후 생성되는 데이터셋의 input, output data를 반대로 하여 instruction tuning 하는 방식으로 학습됨.

sLLM 개발

[콘텐츠 요소 추출 모델 제작용 데이터 가공 과정]



구분	설명
전처리 과정	● 커뮤니케이션북스에서 보유하고 있는 3만개 article 단위의 텍스트 데이터 수집. 위 텍스트 데이터를 맞춤법과 띄어쓰기 수정 등 전처리 과정을 거쳐 사용 가능하도록 가공
데이터 라벨링	● 전처리가 완료된 텍스트 데이터에서 각 article의 콘텐츠 구성요소를 추출. 이 과정은 GPT4와 Human feedback을 이용하여 이루어짐 ● 구체적으로는, 각 article(article에 대한 설명 추가 필요)단위로 콘텐츠의 인물, 장르, 모티프 등 구성요소들을 식별 ● GPT4 사용에는 다양한 방법의 프롬프트 엔지니어링을 통해 위 과정의 정확도를 상승시킴
미세 조정	● 라벨링된 데이터를 sLLM에 미세 조정. input 데이터는 article 단위의 text, output 데이터는 해당 텍스트의 구성요소 라벨 ● 미세 조정을 통해 모델은 텍스트를 입력받아 그 구성요소를 추출할 수 있도록 학습함

- 모델은 선택은 한국어가 가능한 sLLM(Gemma, Polyglot, Solar 등 총 20여 종)들의 양자화, 비양자화 모델 중 학습용 데이터셋을 학습한 후 성능이 가장 뛰어난 모델을 선택.
- 위와 같은 단계를 거쳐 콘텐츠 구성요소를 정확히 탐지하고 분류 가능한 모델을 개발. 이 모델은 추후 생성형 콘텐츠 탐지 및 저작권 분석 기술에 사용됨.

다. 공동연구개발기관 - 오투오

콘텐츠 구성요소 분류 데이터 구축

- 컴복스가 정의한 구성요소 체계를 사용하여 장르, 모티프, 인물 등을 추출하는 방식을 채택. 이를 위해

다음과 같은 단계를 거쳐 데이터베이스를 구축하고 instruction set을 제작함.

- 이와 같은 요소들을 추출하기 위해 30,000만 종 데이터에서 각 요소를 식별하고 라벨링하는 과정을 거쳐야 함. 이를 위해 GPT 모델을 활용하여 데이터를 자동으로 라벨링할 예정임.

과정	구체적인 과정 설명
데이터 준비	• 커뮤니케이션박스의 XML 등 원시자료를 분석에 적합한 구조로 정제 • 맞춤법 수정, 띄어쓰기 교정 등의 전처리 과정을 거침
Instruction 작성	• 구성요소들을 정의하는 instruction 작성 • 각 요소에 대한 구체적인 정의와 예시를 포함하여 모델이 정확히 라벨링
GPT 모델을 이용한 라벨링	• 준비된 데이터와 instruction 을 사용하여 GPT 모델을 통해 자동 라벨링 수행 • 모델이 각 텍스트 데이터에서 해당 요소를 정확하게 식별하고 라벨을 부여하도록 학습
라벨링 데이터 검증 및 수정	• 라벨링된 데이터를 검증하여 정확성을 평가함 • 필요 시 수동 검토를 통해 잘못된 라벨을 수정하고 모델의 성능을 개선함

해당 방식의 데이터는 추후에 표절 탐지 기술에서 요소별 유사도를 판단하는데 있어 비교해야할 요소 선별 과정에서 유용하게 쓰일 것.

컨텐츠구성요소 추출 인스트럭션 데이터셋 설계

- 목적: 콘텐츠 구성 요소의 구체성을 포함하고 있는 정보로서, 콘텐츠 요소 2차 창작물 생성가능.

추출 방식	input data	output
기존 NER	어린왕자 소설의 내용	{ “ 등장인물 ” : [“ 어린왕자 ” , “ 비행기 조종사 ” , “ 여우 ” , ...] }
요소 추출 기술	프롬프트 엔지니어링이 적용된 input data : “ 본문 ” 에 대한 “ 요청 ” 수행, “ 예시 ” 를 참고할 것. “ 본문 ” : <<어린왕자 소설의 내용>> “ 요청 ” : 인물적 요소 추출, depth 2의 json 포맷으로 작성. json내부에는 다음과 같은 변수들이 작성됨 : “ name ” , “ traits ” , “ description ” . “ traits ” 는 인물의 성격을 표현하는 요소 추가. “ 예시 ” 빨간머리 앤, 앤 설리: { “characters”: [{ “name”: “앤 설리”, “traits”: { “imaginative”: true, “talkative”: true, “optimistic”: true, “kind-hearted”: true, “determined”: true }, “description”: “앤 설리는 고아로 자라난 상상력이 풍부하고 말이 많은 소녀로,....” }] }	<pre>{ "characters": [{ "name": "어린왕자", "traits": { "curious": true, "innocent": true, "thoughtful": true, "adventurous": true }, "description": "어린왕자는 순수하고 호기심이 많은 소년으로, 자신의 작은 별에서 여러 다른 행성으로 여행을 하며 다양한 사람들과의 만남을 통해 성장해 나갑니다." }, { "name": "파일럿", "traits": { "reflective": true, "empathetic": true, "lonely": true, "nostalgic": true }, "description": "어린왕자를 사막에서 만난 파일럿으로, 어른이 되면서 잃어버린 어린 시절의 순수함을 어린왕자와의 만남을 통해 되찾게 됩니다." }, ...] }</pre>

스토리 요약/분석 기술

- 정교하고 정확한 콘텐츠의 스토리 요약 제공가능하여, 생성형 콘텐츠 탐지 및 저작권 분석 기술에 활용.

방안	목표 세부 설명
추출적 요약	구성 요소 추출 모델을 활용하여 텍스트 내의 주요 요소들을 식별하고, 이를 기반으로 핵심 내용을 추출하여 요약본 생성
추상적 요약	- 원본 텍스트의 의미를 유지하면서도 간결하고 읽기 쉬운 요약을 생성. 단순히 원문을 축약하는 것이 아니라, 내용의 재구성 및 표현을 포함 - GPT 등의 고성능 텍스트 생성 모델을 활용
서비스 모듈	흐름: 입력 ⇒ 추출적 요약 (중요 정보 식별) ⇒ 추상적 요약(내용재구성) ⇒ 최종 요약

표절 검출 기술

- 1차년도에서 개발된 콘텐츠 구성요소 추출 기술을 기반으로 하여 표절 검토 기술을 개발함. 2차년도 표절 검출 기술 개발은 다음 단계로 진행됨

기능 단계	목표 세부 설명
요소 추출 기술	요소 추출모델을 활용하여 텍스트 내의 주요 인물, 장르, 모티프 등의 구성요소를 추출.
기본 표절 검출 알고리즘	- 분석 대상: 추출된 구성요소 - 분석 방법: ① 텍스트 간의 유사성 비교, ② 구성요소 간 중복되는 내용이나 유사한 표현 분석, ③ 텍스트 매칭과 문맥 의미 분석을 통한 정교화
통합	- 정의: 다양한 콘텐츠에서 구성요소를 추출하고 이를 비교하여 표절 가능성을 평가함. - 서비스모듈화: 입력 ⇒ 요소 추출 ⇒ 알고리즘 적용 ⇒ 결과 출력

ㄴ. 공동연구개발기관 - 바이칼AI

2차 창작용 원천자료 DB 구축 [③ 콘텐츠 생성 지원 DB]

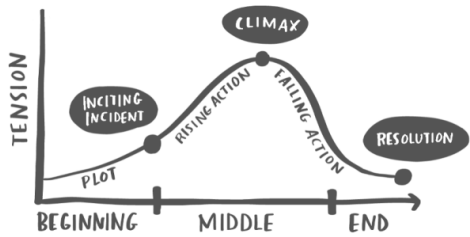
구성	설명
대상 엔티티	- 콘텐츠, 제품, 아티클로 분할하여 구성: 콘텐츠(계약과 연결) - 제품: 종이책, 큰글자책, 전자책 등 동일 콘텐츠가 다양한 포맷과 사양으로 판매가능. - 아티클: 콘텐츠를 작품/주제 단위로 분할한 하위 상품 단위.
장르별 세부 속성 활용	- 도서 기본 서지 정보와 메타 데이터: 200 여개 필드 구성 (도서명, 저자명, 발행일 등 서지정보와 장르, 분류, 간단소개, 세부 소개, 키워드, 편집자 의견 등) - 저작권 및 저작권 침해요소 중 콘텐츠 구성요소 이외의 데이터 구축. 1단계 저작권 침해 데이터 정의에 따라 CCL과 저작권 계약 22항에 근거하여 침해 여부 정의.
논리 DB 설계	- PostgreSQL v15 이상으로 구축 - 메타데이터 JSON 필드로 구축하여 처리(참고: 추천시스템은 neo4j 그래프 DB 사용)

구성	설명
구축 방안	1. DB 설계 후 관리 도구를 구현 2. 데이터 조회 API를 설계하여 API 구성으로 데이터 입력에 대한 검증 3. 데이터 정합성을 샘플링 하여 오류 데이터 발생하지 않도록 검증

컨텍스트 인지 기반 2차 창작을 위한 요소 정의

- 구성요소를 정의하는 것은 AI가 학습하고 이를 통해서 재현해내는 것이 중요함.

구성요소	정의	설명 또는 예시
장르 (genre)	작품의 형식, 내용, 주제, 분위기 등을 특징	- 보편적 장르: 추리, 스릴러, SF, 판타지, 무협, 연애/로맨스, 전쟁, 모험, 역사물, 웨스턴 등 - 한국 고전소설: 전기소설, 환몽소설, 몽유록, 영웅군담소설 등 - 현대적인 장르: 포스트 아포칼립스, 회귀물, 빙의물, 환생물 등
모티프 (motif)	인물과 사건을 동기화시키는 요소	- 장르 문법에 따라, 세부적인 모티프(motif) 옵션을 설정하는 것이 필요함. - 모티프는 작품의 핵심 주제를 이루는 한 부분이면서 인물과 사건을 동기화시키는 요소를 가리킴. - 인물(character), 상황(situation) 등을 조합하여 모티프로 분류하는 경우도 있음. (이화여대 스토리헬퍼의 경우)
문체 (style)	‘장르’ 설정에 영향을 받는 요소	- 문체 혹은 서술 방식 설정은 다양한 방식으로 구분 및 범주화 가능 - 장르별 문체 설정: - 시점: 1인칭, 3인칭 등 - 서술방식: 사건 설명 중심, 인물 중심, 감정적 서술, 논리적 서술
세계관 fictional universe	이야기의 시간적, 공간적, 사회적 배경	- 서사학에서는 서사세계(narrative world) 혹은 스토리세계(storyworld)라고 지칭됨. 이야기의 맥락을 구성하는 요소임. - 유저가 프롬프트를 이용해 별도로 마련된 (가칭) ‘작가 메모란’에 기록하여 개략적인 세계관을 구축함. (시대적) 배경을 포함할 수 있음.
인물(character)	작품 속에 등장하는 사람 혹은 인격체	- 역할: 주인공, 반동인물(antagonist), 적대자, 조력자, 부차적 인물 등 - 속성: 이름, 나이, 직업, 성격, 과거사, 인물 간의 관계 등의 정보 - 인물에 대한 정보는 효과적인 이벤트/액션을 발생시키는 중요한 요소가 됨.
이벤트/액션(event/action)	인물이 벌이는 사건 또는 행동	- 실제로 일어난 일, 즉 시간 순으로 전개되는 사건의 나열을 뜻함 - 추적, 속죄, 미지의 세계로의 여행, 사랑의 힘으로 변신하는 야수, 수수께끼의

구성요소	정의	설명 또는 예시
tion)		해결, 피억압자의 저항, 이방인 구원자, 희망
스토리 아크(story arc)	이야기를 구성하는 일련의 사건의 전개 구조	- 원호(arc) 모양을 가진 이야기의 뼈대로서 플롯에 인과성과 역동성을 부여하고 극적인 서사 패턴을 만들어주는 설계도 역할을 함. 내러티브 아크(narrative arc)라고도 지칭함. - 기본적 스토리 아크: ‘발단 → 전개 → 위기 → 절정 → 결말’의 구조 - 영미권에서 스토리 아크를 설명하는 용어는 ① 설명(Exposition: 등장인물과 배경 소개) → ②사건의 발단(inciting incident) → ③행동 개시(rising action) → ④절정(climax) → ⑤(falling action: 절정 이후의 하강 단계) → ⑥해결(resolution)임. 

“나의이야기” 구성요소에서 적용가능한 구성요소 예시

“나의 이야기”과 관련해서 기존의 “자서전.net”에서 서비스되는 경우, 주제 에피소드 단위로 기존 등장인물, 사건의 유형, 세계관을 선택하여 자동으로 AI 저작을 지원할 수도 있음. (예시적 변

구성요소	정의
장르	- 연애/로맨스, 모험, 역사물, 전기, 에세이 등 - 형식: 일기, 편지, 사진 2차 저작물은 다양하게 변형 가능
모티프	- 연애/로맨스, 전기, 모험 에서 파생된 모티프 추출 가능 - 각 에피소드별로 다양한 모티프 추출 가능
문체	1인칭 시점
세계관	나의 성장기의 시대적, 공간적 배경
인물	- 나: ‘나’는 어떤 사람인가: 성격, 취향, 꿈, 정체성, 가치관 등 - 주변인물: 인간 관계: 가족/친척, 친구, 동료, 연인, 소중한 인연, 은인, 악연 등
이벤트/액션	- 생애 주기별 주요 사건: 출생기, 성장기, 연애, 취업, 결혼, 직장생활 등 - 인생의 중요한 사건: 개인적 사건, 시대적(역사적) 사건, 사건의 성격(희노애락)
스토리 아크	각 에피소드를 묶어서 하나의 스토리 아크로 구성 가능

2차 창작을 위한 컨텍스트 인지 기반 저작도구 (메타정보, 엔티티, 요소 간)

구분	설명
주요기능	- 컨텐츠 구성요소 6(7)가지를 선택할 수 있음. - 구성요소에 따른 세부 구성요소 간의 유기적인 관계를 통해서 만들어짐. - 이를 통해서 자동으로 프롬프트 생성.
프롬프트 예시 (실행결과 만들어지는	CO-STAR Prompt Design **Context:** 마법과 용이 존재하는 중세 판타지 세계에 사는 중년의 사람. 이 인물은 용을 길들이는 마법사입니다.

구분	설명
프롬프트)	**Outcome:** 왕국을 용의 공격에서 구하는 것. **Scale:** 이야기는 작은 야생 생물과의 작은 만남으로 시작하여 점차 더 큰 위협에 직면하는 여정을 포함합니다. **Time:** 전체 모험은 6개월에 걸쳐 이루어집니다. **Actor:** 마법사 아르테미스와 왕국의 공주. **Resources:** 고대 마법 두루마리, 왕실 무기고, 충성스러운 동료들의 지원을 받을 수 있습니다.
샘플 UI 구성	텍스트 엘레나 왕국의 왕녀이다, 용의 습격으로 왕국이 폐멸상태에 이른 지금, 왕녀인 엘레나는 용을 물리치기위해 아르테미스와 모험을 시작하는데.. 인물가이드북 아르테미스 동료 아르테미스, 길거리 용병 주요 사건 용의 공격, 마법사와 왕녀의 모험, 최종 전투, 스토리 아크 선택 이벤트 및 행동 입력 왕녀는 모험을 떠난다. 극적 효과 역동전, 반전이 있는, 열린 결말 작가예모 판타지 적용 결과 등장인물 소개 마법사 아르테미스: 중년의 용을 길들이는 마법사로, 지혜롭고 경험이 풍부하다. 오랜 시간 왕국을 지키기 위해 애써왔다. 공주 엘레나: 왕국의 공주로, 용의 공격으로부터 왕국을 지키기 위해 아르테미스와 함께 모험을 떠난다. 용감하고 결단력이 강하다. 분석 장르 중세 판타지 사건 설명 중심 1인칭 주인공 시점 이야기 배경 마법과 용이 존재하는 중세 판타지 세계 주요 모티프 마법, 용, 왕국 주인공 아르테미스 주인공 설명 용을 길들이는 마법사

메타 정보 기반 교정·교열 기술 개발

구분	설명
필요성 및 기능	[필요성] 1차 창작물: 신규 콘텐츠 구성요소를 추출하여 “스토리 전달의 효과성” 배가 2차 창작물: 변경된 구성요소(모티프,문체,인물,이벤트 등)를 반영하여 “컨텍스트 인지하여” 전달 [목적] - 각 교정 포인트를 중점적으로 확인하는 데 사용 - 이를 통해 글의 일관성과 자연스러운 흐름을 유지하면서 독자의 공감을 유인 가능.
교정의 구현	GPT 프롬프트를 기반으로 sLLM에 관련 학습 추가 구현하여 진행
교정의 예시	I have made significant changes to the characters in a short article consisting of 2 paragraphs and involving a love triangle. Please assist in proofreading and editing the revised version to ensure coherence and consistency. Below are the details of the original and the revised works: **Original Work** - **Title**: "At the Crossroads of Love and Friendship" - **Story**: <<< 중간생략 >>> **Revised Work** - **Title**: "The Triangle of Fate" - **Story**: <<< 중간생략 >>> Please proofread and edit the revised work, focusing on the following correction points: - Ensure character traits and actions are consistent. - Match the narrative tone and style to the updated character traits. - Maintain the logical flow of the story with the revised characters. - Accurately express the emotions of the characters in line with their new traits. - Ensure the dialogues between characters are natural and fitting. - Evoke empathy from the readers through well-depicted internal conflicts and emotions.

컨텍스트 인지 기반 2차창작을 위한 요소 추천 알고리즘 및 구현

구분	세부내용			
추천의 흐름	① 사용자의 입력 분석 - 사용자가 작성한 자서전 내용을 분석하여 주요 주제, 인물의 성격, 중요한 사건 등을 추출 - 자서전의 현재 상태와 방향을 파악하여 필요한 개선점이나 보완할 부분을 인식 ② 스토리텔링 데이터베이스에서 검색 - 구축된 콘텐츠구성요소 DB에서 - 각 요소를 태그와 함께 저장하여 쉽게 검색하고 추천가능 ③ 추천 알고리즘 적용 3개 적용: 콘텐츠 기반 필터링, 협업 필터링, 혼합 방식 사용			
알고리즘	알고리즘	동작	장점	단점
	콘텐츠 기반 필터링	- 텍스트 분석을 통해 사용자의 자서전에서 핵심 키워드와 주제 추출 - DB에서 비슷한 키워드와 주제를 가진 전기의 요소를 찾아 추천	- 사용자 맞춤형 - 특정 주제나 스타일별 정교	- 사용자 데이터 부족시 적용 불가 - 선호도 변화시 적응 불가
	협업 필터링	- 유사한 사용자가 자주 선택한 요소 추천 - 사용자 간의 유사성을 계산하여 추천 리스트 생성	- 넓은 범위의 추천 - 빠른 요소 검색	- 사용자 데이터 부족시 적용 불가 - 유사성 속도 계산지연
	혼합 방식	사용자 입력과 다른 사용자의 데이터를 동시에 활용하여 추천	- 정확성, 다양성 - 선호도 변화에 대응	- 구현 복잡도 증가 - 많은 리소스 필요
DB설계	- neo4j 그래프 DB를 사용하여 설계 - 노드와 관계 설계: 사용자, 아티클, 키워드, 각각의 구성요소 (:HAS_KEYWORD), (:SIMILAR_TO) 관계 // 유사 사용자 기반 추천 예시 MATCH (u:User)-[:SIMILAR_TO]-(su:User)-[:RECOMMENDED]->(b:Biography) WHERE u.id = \$user_id RETURN b.id, COUNT(*) AS recommendation_score ORDER BY recommendation_score DESC LIMIT 10;			

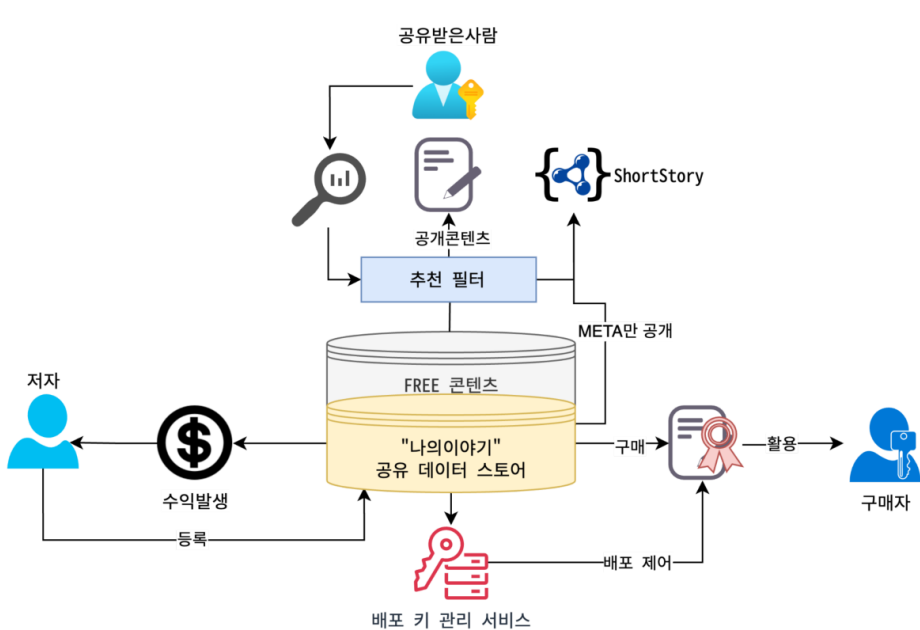
공유서비스 개발

서비스 개요

구분	세부내용
주요기능	전체 서비스는 ①공유, ② 콘텐츠 구매, ③ 저작지원으로 구성 공유서비스는 카테고리 / 커뮤니티 / 마켓으로 구성 - 카테고리: 작가별, 장르별, 주제별, 문제별, 인물별, 이벤트별로 구성하여 검색 - 커뮤니티: 카테고리별로 작가와 사용자가 소통하는 공간, 사용자간 소통, SNS를 통한 공유 - 마켓: 마켓에서 카테고리별로 콘텐츠 추천, 구매, 유통 - 지원: 저작 지원, 출판 지원으로 구성
사용성	(사용자 접근방법) 모바일 앱 및 웹을 통해서 모두 사용 가능 (알림서비스) 앱 PUSH 및 웹 푸시 기능을 구현하여 접근성 개선

구분	세부내용
	• (카카오톡 알림) 카카오톡 알림 서비스를 통해서 웹 및 앱 링크 제공 • (사용자층에 대한 고려): 10대부터 70대까지 사용할 수 있는 서비스 ◦ 시각적요소: 색상대비, 가변적 폰트크기, 명료한 텍스트 사용 ◦ 청각적요소 활용: 음량 조절, 청각적 반응을 제시 ◦ 최소 정보 표출: 단일 정보에 최소 정보 표출 ◦ 연령대별 각 10명의 UX 테스트를 거쳐서 다양한 사용자 요구사항 반영
추천	• 나의 이야기 작성에 도움이 되는 콘텐츠 추천 • 2차 창작을 위한 콘텐츠 추천
저작 지원	1차 저작 및 2차 저작을 위한 콘텐츠 구성 및 배치를 통한 새로운 콘텐츠 작성 가이드
공유	API 방식을 통해서 페이스북, 인스타그램, 트위터, 링크드인, 유튜브, 카카오톡, 라인 등 지원

공유 서비스 아키텍처 설계

구분	세부내용
클라우드	• KT 클라우드 (GiGA Stealth) 또는 NCP (Global Edge) 채택 • 클라우드 스토리지와 데이터베이스를 연동하여 콘텐츠 저장 및 관리
아키텍처	• net4j 그래프 DB 사용하여, 노드 {사용자}, {아티클}, {콘텐츠구성요소} 관계 표시 • Elasticsearch를 활용한 고속 검색 기능 구현 및 인덱싱 • OAuth2 기반 사용자 인증 및 권한 관리 시스템 구현, Google Analytics:사용자 행동 분석
서비스 흐름	• 공개 데이터와 상용 데이터 구분 • 상용 데이터에 대한 키관리를 통한 배포, AES256-CBC 암호화 적용 & 접근제어 • 카카오톡 등 각종 인증 서비스 연동  다이어그램 설명: 공유받은사람(사용자)은 공개콘텐츠를 추천 필터를 통해 '나의이야기' 공유 데이터 스토어에 업로드합니다. 이 스토어는 'FREE 콘텐츠'와 'META만 공개' 콘텐츠를 포함합니다. '나의이야기' 공유 데이터 스토어는 '수익발생'을 통해 '저자'에게 '등록'됩니다. '나의이야기' 공유 데이터 스토어는 '구매'를 통해 '구매자'에게 '활용'됩니다. '구매자'는 '배포 키 관리 서비스'를 통해 '배포 제어'를 받습니다. 'ShortStory'는 '나의이야기' 공유 데이터 스토어와 연결되어 있습니다.

구분	세부내용
공유 기술	- 콘텐츠별 고유 링크 생성 ⇒ schema.org의 Article, ShortStory, Author 객체와 연동 - JSON LD 구성 ⇒ 구글 검색 등에서 사용이 가능, 권한에 따라서 부분 공개 가능. <pre>{ "@context": "https://schema.org", "@type": "ShortStory", "@id": "https://nutory.me/shortstory/gih2yun/unique-story-id", "author": { "@type": "Person", "name": "gih2yun"}, "name": "The Unique Tale of Adventure", "description": "A thrilling adventure story that takes the reader on an unforgettable journey.", "datePublished": "2024-05-22", "genre": "Adventure", "keywords": ["adventure", "journey", "thrilling", "unique"], "text": "Once upon a time in a land far, far away...", "publisher": { "@type": "Organization", "name": "Nutory"}, "url": "https://nutory.me/shortstory/gih2yun/unique-story-id" }</pre>

나-1. 2단계

모델	▶ 요소분석 및 추출, 교정/교열, 추천, 2차 창작 모델 고도화 관계형 행동 정의에 따른 저작 기술 추가
데이터	▶ 저작권 침해요소에 대한 식별 및 케이스 분석 DB
서비스	▶ 저작권 침해요소 및 저작기술 고도화를 반영한 콘텐츠 공유 플랫폼 출시

나-2. 2단계 - 개발목표

유형	기관	개발목표
주관	커뮤니케이션북스	③ 콘텐츠 생성 지원 DB 구축 - 저작권 침해요소별 메타 데이터 분석 사례 리포트 - 침해요소 케이스 아카이빙
공동1	고려대학교산학협력단	① 출판언어모델 / SCI급 논문(1건게제), 생성모델 고도화
공동2	오투오	② 요약 기술 개발 / 표절 검출 기술 개발, 추상적, 추출적 요약 기술 개발, 통합 요약 시스템 개발, 통합 표절 검출 시스템 개발
공동3	바이칼에이아이	③ 콘텐츠 생성 지원 DB 구축 / DB 공유 기술 ④ 저작기술 / 관계형 동작 기반 저작기술 개발, 추천 알고리즘 고도화 ⑤ 콘텐츠 제공&공유기술 / 저작권 침해 정보 케이스 분석, 공유서비스 고도화

나-3. 2단계 - 개발내용

ㄱ. 주관연구개발기관 - 커뮤니케이션북스

저작권 침해요소별 메타데이터 분석 사례 리포트 [③ 콘텐츠 생성 지원 DB 구축]

- 1 단계에서 정의한 ‘저작권 침해 데이터 정의’에 따라 저작도구1~3에서 창작한 콘텐츠의 표절, 저작권 침해 여부를 점검함. 점검 결과 탈락하는 콘텐츠에 대한 분석 실행.
- 침해 요소는 ①CCL, ②저작권 계약서의 권한 항목에 준해 정의함.

침해요소 케이스 아카이빙 [③ 콘텐츠 생성 지원 DB 구축]

- 1단계에서 저작도구 1~3을 개발하여 운영. 운영 과정에서 발생하는 다양한 침해 사례에 대해 분석을 실시하고, 그 결과를 아카이빙함. 법적, 정보를 포함하여, 메타데이터 등을 추가하여 확보 해야 함.
- 즉, 침해 창작물과 해당하는 침해 요소를 함께 메타정보를 포함하고 법적기간, 접근권한을 고려하여 DB화.

ㄴ. 공동연구개발기관 - 고려대학교 산학협력단

요소파악 기술 (고도화)

구분	내용
베이스모델	콘텐츠 구성요소 추출 모델 sLLM
요소 파악 기술 고도화	• 기존 LLM 기본 기능에 대한 성능 평가를 통해서 • (알고리즘 개선) 모델의 하이퍼파라미터를 조정하고, 새로운 피처를 추가 • (데이터셋 확장) 다양한 도메인과 상황에 맞는 데이터를 추가하여 모델의 범용성을 높임. • (검증 및 최적화) 최적화된 모델을 선정하고, 성능을 지속적으로 모니터링함

도메인 장르별 AI 창작 기술 개발

구분	내용
베이스모델	2차 창작 생성지원 LLM
구현 방법	• 도입기술: ① Prefix Tuning, instruction tuning 등의 기술, ② Prompt Engineering

LLM 기반 저작권 침해 방지 기술

- 제약적 제어 가능 생성 알고리즘(조건부 생성 모델, 규칙 기반 제약)을 적용하고 추가학습 진행.
- 생성 디코딩 알고리즘에 저작권 침해 방지 기술을 통합하여 저작권 침해 방지 모듈을 제작

방법	세부 내용
저작권 침해 감지 알고리즘 통합	• (실시간 탐지) LLM에 저작권 침해를 실시간으로 감지할 수 있는 알고리즘을 통합함 • (유사성 분석) 모델이 텍스트를 생성하는 과정에서 기존 저작물과의 유사성을 분석하고, 저작권 침해 가능성을 사전에 차단함 • 기존 저작물과의 유사성 분석은, 1단계에서 제작한 요소 추출 모델을 이용하여 글의 요소를 파악하고, 유사한 요소간 유사도를 비교하는 이중 유사도 분석 알고리즘을 통해 분석함
제약적 생성 알고리즘 적용	• Constrained and Controllable Decoding Strategies를 통해 생성 과정에서 저작권 침해를 방지함 • 생성 텍스트의 특정 패턴이나 표현이 기존 저작물과 중복되지 않도록 제어함
추가 학습을 통한 환각 현상 방지	• LLM의 환각 현상을 줄이기 위해 추가적인 학습 데이터를 활용하여 모델을 개선함 • 저작권 침해 가능성을 낮추기 위해 법적 기준에 맞는 데이터셋을 구축하고 이를 통해 모델을 재학습함

㉔. 공동연구개발기관 - 오투오

스토리 요약/분석 기술 고도화

과정	세부 내용
데이터 수집 및 전처리	- 다양한 분야의 텍스트 데이터를 사용하여 요약 모델의 범용성을 높임 - 각 도메인의 특성을 반영할 수 있도록 도메인별 데이터셋 구축
고도화된 요약 모델 개발	- 기존의 문장 추출 방식을 개선, 문맥을 고려한 중요한 문장 추출 기법 개발 - 텍스트 내 키워드 및 문장 간 관계를 분석하여 보다 정밀한 요약 제공 - GPT-4 와 같은 언어 모델을 활용하여 자연스러운 문장 생성 강화
통합 요약 시스템 개발	- 추출적 요약과 추상적 요약을 결합한 하이브리드 요약 모델 개발 - 원문에서 중요한 문장을 추출한 후, 이를 기반으로 자연스러운 요약문을 생성하는 방식으로 정확성과 자연스러움을 동시에 달성함
사용자 맞춤형 요약	사용자 설정 옵션 추가: 요약 길이, 상세도, 강조할 내용 등을 사용자가 설정할 수 있는 옵션을 제공

표절 검출 기술 고도화

구분	내용
고도화 전략	(군집화) 분류된 요소가 비슷한 텍스트들을 군집화하여 해당 군집 내의 요소간의 유사도를 비교하는 방식으로 표절 기술을 고도화. 기존의 표절 검출 기술과 함께 사용하여 단순한 복사하기식 표절은 쉽게 판별이 가능하며, 일부 등장 인물만 바꾸거나 특정 요소만 바꾼 표절글도 표절여부의 수치화가 가능 (Human Feedback) Human Feedback 형태의 Preference Optimization을 통해서 표절 문서를 비선호하게끔 학습하는 형태로 sLLM을 고도화.
점검방법	표절 판별 여부 자체는 자체 구축된 데이터 셋에 테스트하여 판별 수치를 정의 가능.

㉕. 공동연구개발기관 - 바이칼AI

요소 간 관계형 동작 정의 기반 저작 기술 개발

개념: 컨텍스트 인지 기반 2차 창작

- 2차 창작: 원저작물을 이용해서 인용, 인물대체, 대체 플롯, 배경 변경 등을 수행
- 컨텍스트 인지: 기존 구성요소 간의 맥락을 인지한 상태에서 새로운 변형을 가하는 방식, 이 방식은 인물, 사건, 배경(세계관), 스토리 아크 등의 변화를 수반하는 경우임.

구분	세부 유형	예시
배경설정	새로운 배경 설정	원작에서 묘사된 세계의 다른 부분을 탐험하는 이야기를 창작
	배경 확장	원작의 배경 설정을 더욱 깊이 있게 묘사하거나, 다른 장소를 추가할
스토리 아크	플롯 확장	원작의 이야기를 이어가거나, 중간에 새로운 사건을 추가하여 이야기를 확장
	대체 플롯	원작의 플롯을 변형하거나 다른 방향으로 전개하는 대체 이야기를 창작
인물	새로운 인물	원작의 캐릭터와 유사한 성격이나 배경을 가진 새로운 캐릭터를 창조

구분	세부 유형	예시
이벤트	사건의 변형	한 새로운 모험,

컨텍스트 인지 기반 2차 창작을 위한 추천 알고리즘 고도화

- 기존 추천 알고리즘을 구성하는 추천 요소에 “인물, 사건, 배경의 연관짓는” 공통 요소인 모티프를 연계하여 구성
- 모티프는 작품의 핵심 주제를 이루는 한 부분이면서 인물과 사건을 동기화시키는 요소를 가리킴.
- 모티프는 아티클 단위로 나타날 수 있음. 모티프는 대략 **207개로 정의**할 수 있음(2023,김지선).

관계형 동작 정의

- 관계형 동작 정의는 인물과 사건의 관계인 모티프를 변경, 생성시키는 과정. (아래 예시 참조)

주체	관계 정의	객체	예시
인물	관계 변형	인물	인물(주체, 객체)의 친소, 상하, 사제 등의 관계, 강도, 적대성 등의 변형
인물	속성 변경	인물	주체(인물) 또는 객체(인물)의 속성을 변경함으로써 작용하는 힘의 변형
인물	사건 추가	인물	주체(인물)이 객체(인물)에게 변형된 행동 또는 새로운 행동으로 변형
배경	이벤트 추가/변형	인물	배경에서 인물에서 새로운 사건이 발생하는 경우, 또는 기존의 변경
인물	액션 추가	배경	인물이 배경에 대해서 새로운 액션을 추가하는 경우 등

관계형 동작에 따른 저작 기술

구분	세부내용
프롬프트	● 프롬프트: 관계형 동작 유형에 따른 프롬프트를 제시하고, LLM을 통해서 새로운 동작 정의 ● 관계형 동작의 풍부함: LLM 학습시 instruction 셋을 통해서 해당 변형이 주어진 경우에 대한 새로운 2차 창작을 few shot으로 학습하도록 함.
이론적 근거	● 프로프의 민담형태론: 사건과 행위를 31개의 형식으로 추출하여 공식화 ● 유철균의 스토리헬퍼: 207개의 모티프를 인물, 상황에 따라 구성
예시	아래는 원작의 주체가 객체의 사랑의 고백을 받아들이게 바꾸는 경우, 글의 생성을 돕는 프롬프트 예시임. Context: 중세 판타지 세계. 전투가 끝난 후, 왕국의 축제에서 용감한 기사(주체)와 이웃 왕국의 공주(객체)가 만나고 있습니다. 이 둘은 처음에는 적대적이었으나 시간이 지나면서 서로에게 호감을 느끼게 되었습니다. Objective: 축제의 밤에 객체가 주체에게 사랑을 고백하고, 주체가 그 고백을 받아들이는 장면을 작성합니다. Style: 서사적이고 감성적인 판타지 스타일로 작성합니다. Tone: 따뜻하고 로맨틱한 톤을 유지합니다. Audience: 판타지 장르를 좋아하는 독자들, 특히 로맨틱한 요소가 가미된 이야기를 선호하는 청소년과 성인 독자. Response: 텍스트 형식으로 작성합니다.

저작권 침해 자동 분석 모듈

- 저작권 침해를 자동으로 판단하는 AI 또는 규칙기반 기능 개발
- 저작권 구매 등 권한에 따라서 ① 복제 등 1차 저작권 침해 및 ② 변형 등 2차 저작물에 대한 침해 여부 자동 판단 하는 기술을 적용
- 2차 저작물에 대해서도 1차 저작물의 구매 및 허용 범위에 따라서 저작권 적용 범위를 자동으로 계산할 수 있도록 구성하고, 추적성을 보장할 수 있도록 하여, 침해를 자동분석하는 성능을 향상.

저작권 침해 정보 케이스 & 공유서비스 고도화

- 기존 공유 서비스 체계에 [저작권 침해 정보를 분석하는 기능](#)을 부착하여 공유 서비스를 고도화함.
- [저작권 침해 요소 정의와 분석의 결과를 반영하여](#), 시스템을 통해 저작권 침해 요소들을 정리하고 허용 가능한 작업만 이뤄지도록 서비스 고도화
- 저작권 관련 행위 및 제어 기술 ⇒ 행위를 인지 후 권한에 따라 허용. 침해행위 인지 시 사전 차단.

저작권 관련 행위	복사, 붙여넣기	스크린샷	프린트	저장 & 다운로드	임베드	무단 표절
대응기술	클립보드	CSP	Print Detection	다운로드 모니터링	워터마크	표절 탐지

- 2차 저작물에 대해서도 동일하게 저작물로 자동으로 등록되어 저작권의 권리로 자동 관리하도록 구성.

2-3. 연구개발의 창의성 · 혁신성 등

특성	비교 대상	개발 목표의 우수성
추출 모델과 생성 모델의 분할	GPT	sLLM의 성능 구조에서 다양한 모델을 연결하는 방식으로 재구성. 인스트럭션 학습을 과정을 고려해서 임무에 맞게 베이스모델 구분.
창작 유형 별로 저작도구 구분 개발	기존 저작도구	생성형 AI기술 적용하여 창작 유형에 따라 3가지로 구분하여 도구 개발
창작의 문턱을 낮춤	전문출판시스템	다양한 개인의 경험을 창작자산으로 변경할 수 있는 접근 방법으로 “나의 이야기”로부터 출발. 누구나 창작을 할 수 있는 방안을 제시.
schema.org를 통한 창작 콘텐츠 공유	inkitt.com	웹 검색을 통해서 글의 내용 및 메타가 검색되고, 저자, 장르별로 검색이 가능함.

2-4. 연구개발 성능지표 및 평가방법

가-1. 결과물의 성능지표

No	주요 성능지표	단위	개발목표치			비교수준	가중치 (%)	평가방법
			1-1년	1-2년	2-1년			
1	sLLM 콘텐츠 추출모델 환각방지능력 (Ko-TruthfulQA - Accuracy)	%	-	80	85	Hugging face ko LLM Leaderboard 상위 10개 평균 (84.047)	15	공인인증
2	sLLM 2차저작 생성지원 모델 상식생성능력 (Ko-CommonGenV2 - Accuracy)	%	-	50	55	Hugging face ko LLM Leaderboard 상위 10개 평균 (53.095)	15	공인인증
3	메타 데이터 추출 F1	%	-	80	83 ^{이상}	KLUE NER Leaderboard 등재되어 있는 모델 중 top 5 이내	10	공인인증
4	표절 여부 판별 정밀도 (precision)	%	-	95	97		10	공인인증
5	콘텐츠 요소 추천 NDCG	%	-	75	80	-	10	공인인증
6	메타기반 교정/교열 정확도 (word level accuracy)	%	-	85	90	구글, 마이크로소프트의 맞춤법 교정 알고리즘 (92%)	10	공인인증
7	요약 성능 (N-gram Rouge score)	%	-	60	65	gpt 4o의 줄글 요약 스코어(64%)	10	공인인증
8	정량 사용성 평가	점	-	75	85	-	20	외부기관

가-2. 평가방법 및 평가환경

No.	주요 성능지표	세부 평가방법 및 평가환경
1	sLLM (Ko-TruthfulQA-Accuracy)	- Open Ko-LLM 리더보드의 5개 벤치마크 중 하나 - 학습된 언어 모델이 주어진 질의(Question)에 대해 진실된 응답에 가깝게 추론을 생성하는지 혹은 틀린 응답들에 가까운지 판별하게 하는 데이터로 구성 - 언어 모델이 사람의 잘못된 판단을 얼마나 제어할 수 있는 지를 수치화 - 데이터셋에 대한 Accuracy 계산 - 평가 환경 : 하드웨어 : A100 GPU, 소프트웨어 : python version 3.9, pytorch version 2.2.2+cu121, transformers version 4.39.3
2	sLLM (Ko-CommonGen v2-Accuracy)	- Open Ko-LLM 리더보드의 5개 벤치마크 중 하나 - 언어모델이 일상 상식을 얼마나 보유하고 있는지를 평가 - 데이터셋에 대한 Accuracy 계산 - 평가 환경 : 하드웨어 : A100 GPU, 소프트웨어 : python version 3.9, pytorch version 2.2.2+cu121 transformers version 4.39.3
3	메타데이터 추출	- 평가 방식 : KLUE 데이터셋의 NER(개체명 인식) 데이터셋을 활용하여 약11,000개의 학습 데이터로 개발된 모델을 학습시키고, 2000여개의 테스트 데이터를 활용하여 개발된 모델에 대한 상대 평가 진행. 상대 평가 기준으로는 KLUE 데이터셋 Leaderboard에 등재되어 있는 10여개의 baseline들을 기준으로 비교 진행 - 평가 환경 : python 3.9
4	표절 여부 판별 정밀도 (precision)	- 평가 방식 : 자체 제작된 실제 표절 글과, 표절이 아닌 글을 classification 하여 precision 계산. - 평가 환경 : data: 자체 제작된 실제 표절 글 준비, Ubuntu 22.04, python 3.9, 자체 test script
5	콘텐츠 요소 추천 NDCG	- NDCG: normalized Discounted Cumulative Gain - 사용: 상위의 랭킹 리스트가 하위 랭킹 리스트 보다 확연하게 중요한 도메인에서는 유용한 평가 기준 - 평가방식: 모든 테스트 셋 유저들에 대한 추천의 평균 NDCG를 구하여 모델의 성능을 평가, 100명의 테스트 셋 유저들을 대상으로 평가 - 평가 환경 : data: 사용자, 사용자 검색 추천 조건, Ubuntu 22.04, python 3.9, 자체 test script $nDCG_p = \frac{DCG_p}{IDCG_p}$ Where $IDCG_p = \sum_{i=1}^{ REL_p } \frac{2^{rel_i} - 1}{\log_2(i + 1)}$
6	메타기반 교정/교열 정확도	- 평가 방식 : AI hub의 자연어 분석 후처리용 과교정 데이터(200만여 건)를 사용하여 잘못된 문장의 단어 수준 정확도 계산 - 평가 환경 : data: AI hub의 자연어 분석 후처리용 과교정 데이터, Ubuntu 22.04, python 3.9, 자체 test script
7	요약성능	- 평가 방식: 요약 데이터셋을 구축하여 n-gram rouge 스코어를 다음과 같은 수식을 통해 계산함 - 평가 환경 : 하드웨어 : A100 GPU, 소프트웨어 : python version 3.9, pytorch version 2.2.2+cu121 transformers version 4.39.3 $ROUGE - N = \frac{\sum_{S \in \{Reference\ Summaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{Reference\ Summaries\}} \sum_{gram_n \in S} Count(gram_n)}$
8	정량 사용성(UX) 평가	- 평가방법: 사용성 평가를 수행할때, 정량 데이터를 수집하여 평가: ① 사용자의 수행 모습을 직접 관찰하고 측정한 객관적인 데이터 획득, ② 소수의 인원을 대상으로 반복적인 피드백 루프를 수행 - Task Time(수행시간), Errors(오류 횟수), Success Rate(완료율), Efficiency(효율성), Learnability(학습용이성)에 대한 달성률을 계산하여 평가 - 평가 환경 : 하드웨어 : A100 GPU, 소프트웨어 : python version 3.9, pytorch version 2.2.2+cu121 transformers version 4.39.3

나. 성과물 목표

단계	성과물 목표		
1 단계	[개발산출물]		
	컴복스	장르 메타정보 정의서 1식 콘텐츠 구성요소 정의서 1식 학습용 파인튜닝 데이터 3천개 원천 자료 데이터화 3만개 세부 속성정보 DB 1식	데이터품질 (의미정확성 95%)
	고려대	출판콘텐츠추출모델 1개 2차저작 생성지원 모델1개	성능지표 참고
	오투오	콘텐츠 요소 분석 AI 서비스 모듈 1식 콘텐츠 표절 여부 AI 탐지 모듈 1식 SW 저작권 등록 1건	성능지표 참고
	바이칼AI	엔티티&메타정보활용 DB 1식 엔티티 DB관리 관리도구1식 요소 추천 서비스모듈 1식 교정교열 AI 서비스모듈 1식 2차 저작 AI 서비스모듈 1식 콘텐츠 공유 플랫폼 1식 SW 저작권 등록 3건	성능지표 참고
[과학적성과]			
● 비SCI급 논문 3건			
2 단계	[개발산출물]		
	컴복스	저작권 침해요소 아카이빙 1식 콘텐츠공유플랫폼 기획안 1식 mutory.me(가칭) 베타 서비스 오픈 1식 출판용 교정교열 시스템 1건 개발 공정 개선	성능지표 참고 사용성 평가
	고려대	출판콘텐츠추출모델 1개 (고도화) 2차저작 생성지원 모델1개 (고도화)	성능지표 참고
	오투오	콘텐츠 요소 분석 AI 서비스 모듈 1식 (고도화) 콘텐츠 표절 여부 AI 탐지 모듈 1식 (고도화) SW 저작권 등록 1건	성능지표 참고
	바이칼AI	저작권침해요소 분석결과 보고서 1식 요소 추천 서비스모듈 1식 (고도화) 2차 저작 AI 서비스모듈 1식 (고도화) 콘텐츠 공유 플랫폼 1식 (고도화) 저작권 침해요소 케이스 분석DB 1식 SW 저작권 등록 2건	성능지표 참고
[과학적성과]			
● SCI급 논문1건 ● 특허 2건 출원 (스마트지수 B등급 1건 포함)			
과제 종료 후 (2년 이내)	[경제적성과]		
	● mutory.me 서비스 확대에 따른 개인 사용자 1만명 달성 ● (단체 사용자) 출판사 3개 공동 참여, 기관 100 공동 참여 달성 ● 유료 등록 콘텐츠 5만건 달성, 2차저작물 10만건 달성		

3. 연구개발의 추진전략·방법 및 추진일정

3-1. 연구개발 추진전략·방법

경험, 신뢰, 실력을 갖춘 컨소시엄

- 35년 업력 7개 브랜드, 저자 3700명, 5.1만 종 보유
- 신규 출판 사업 경험 풍부
- TIPa, KEIT, NIA 등 다수의 R&D과제 주관 수행
- 엔터프라이즈 생성형 AI 솔루션 연구 및 납품 실적
- 다양한 도메인 메타 데이터 추출 및 응용 서비스 실적 보유



- NLP / 음성분석 AI 기술에 특화
- 상용서비스 개발 및 제품화 실적 보유
- 검증된 연구개발 수행 능력
 - NIA, AI학습데이터 사업 2년연속 우수기관 선정(2건)
- 고려대 Human -inspired AI 연구소
- 5명의 교수와 30여명의 연구진
- 자연어처리, 인공지능, 딥러닝 연구
- 활용 가능한 생성 AI 언어모델 보유
 - 자체 개발 KULLM

전문가 그룹의 자문 활용

- 스토리텔링, 저작권, 자서전(학습생애사, 나의 이야기), 데이터베이스, AI기술 등 다양한 전문가 투입.

영역	이름	직함	소속
문학, 스토리텔링	천정	교수	성균관대학교 국문학과
출판, 스토리텔링	이용희	박사	성균관대학교 국문학과
학습생애사, 평생교육	강대중	교수	서울대학교 교육학과, (전)국가평생교육진흥원장
전 생애발달, 평생교육	이지혜	교수	한림대학교 일송자유교양대학(교육학과)
국어학, 구문론	박진호	교수	서울대학교 국어국문학과 교수. 형태소분석엔진 개발
저작권법	우경선	변호사	법무법인 자연

장애 요소 및 해결 방안

구분	난제 및 장애 요소	해결방안
저작권	AI학습 데이터의 저작권 침해	• 정식으로 공개된 데이터와 저작권을 해결한 학습용 데이터만 사용 • 지금까지 연구결과와 데이터 바탕으로, 사업화에 초점을 맞추어 파인튜닝 데이터 세부 설계하여 적용 • “지식을만드는지식” 브랜드 보유, 국내외 고전문학 한국소설문학선집 100권, 한국수필문학선집 40권, 지만지드라마의 희곡 500여 편 등에서 선별하여 활용.
	생성형AI를 활용한 창작물의 저작권	• 생성형AI 엔진이 적용된 저작도구로 창작과정에서 창작자의 사상과 의도가 자연스럽게 적용되도록 저작도구와 활용 DB를 설계함. 저작도구를

구분	난제 및 장애 요소	해결방안
	획득	사용하는 과정에 자연스럽게 저작권이 발생함.
	플랫폼 활용 창작물의 저작권 침해 방지	- 콘텐츠 재료는 모두 저작권이 확보된 것을 사용함. - 창작물을 저작권 계약할 경우 저작권 침해 여부를 자체 개발한 도구를 활용하여 검사함.
	저작권 추적과 수익배분	- 저자와 계약관리용 자체 “★자산관리시스템2”를 개발하여 모든 콘텐츠▶제품▶아티클 단위에 제품 식별코드를 부여하고, 키 배포 관리를 통해 저작권 허락 범위에 따라 활용될 수 있도록 플랫폼 설계 관리함. - 블록체인 기반한 NFT 적용 검토하였으나, 플랫폼이 기술 배타성을 갖게 되어 대중 접근성이 떨어짐. 일반적인 DB 기반 자산관리시스템을 바탕으로 호환성 확보, 활용성 높임.
사업성	창작 콘텐츠 품질 확보	- 출판사가 보유하고 있는 대중성 있는 고전 문학 중심으로 기본 콘텐츠 데이터를 구축하여 보유 콘텐츠DB의 품질을 확보. - 오픈 데이터 가운데 스토리 생성에 부합하는 주요 데이터를 바탕으로 장르와 학습 공정 설계하여 학습효과와 비용 효율성 확보
	이용자 접근성과 매출 기회 확보	- 당사의 미디어 익스텐션 전략에 따라 동일한 콘텐츠를 다양한 파생상품(종이책, 전자책, 큰글자책, 아티클)으로 동시 출간하여 이용자의 접근 채널 다각화, 매출 기회 확보. - 특히 디지털 콘텐츠뿐 아니라, 종수에 상관없이 종이책 형태로도 판매. 주관사의 무재고 주문생산 기술에 따라 1부도 제작 유통 가능. - 생성형AI로 제작한 콘텐츠 퀄리티가 우수할 경우 출판사 판단으로 정식 출간함. ISBN, 납본과 서점 판매 등 정식 유통.
기술적 측면	콘텐츠 구성요소 추출 정확도 문제	- 콘텐츠 구성요소 정확도를 높이는 것이 본 과제의 핵심적 기술 문제임. 인스트럭션 데이터셋을 GTP로 구축하고, 전문가 검토를 통해서 정확도 향상, 인물, 사건, 세계관 등에 전문가(교수진)를 투입하여 정확도 향상. - 현재 자서전.net 사용자들의 피드백을 통한 정확도 향상 방안 제시.
	창작물의 다양성 및 2차 창작의 효과성	- 콘텐츠 구성요소에서 각 구성 요소 간의 차이를 벡터로 인지해내는 기술이 필요함. ① 콘텐츠 구성요소 원시 데이터의 다양성을 확대하여 문제 해결. ② 유사벡터와 차이벡터에 대한 클러스터링 기법으로 유효 성능 확보. - 영향받는 응용 모델: 교정/교열, 2차 창작 저술, 관계형 동작 정의 기능 등
	표절 검출 정확도	콘텐츠 요소 추출 구체화, Human Feedback, 비교 방법 다양화 적용.
	콘텐츠 활용 및 제어	콘텐츠 공유 플랫폼 내 사용자의 행위 탐지하여 권한에 따라 제어.

3-2. 연구개발 추진체계

기관명	담당 연구개발내용	연구개발 역량의 우수성
주관연구개발기관 (커뮤니케이션박스)	- 장르 선정과 학습용 데이터 확보 - 콘텐츠 구성요소 정의와 특성 분석 - 저작권 침해 데이터 정의, 사례와 케이스 아카이빙 - 공유 서비스 모델을 위한 DB 구축 - mutory.me 서비스 사업 기획 운영	35년 업력 7개 브랜드 보유 출판사. 집필진 3700명 출판 네트워크와 5.1만 종 출판자산 보유. 무재고 주문생산 기술 보유(2015 벤처기업, 기보 인증) - 리딩패킷 서비스(2009~), 큰글자책(2013~), 오디오북(2014~) 등 출판 포맷 다양화, 신규 출판 비즈니스 개발 운영 경험 풍부. - 구술자서전출판서비스(TIPA), 문해한국어서비스(KEIT),

		인공지능데이터구축사업(NIA) 등 다수의 R&D과제 주관 수행.
공동연구개발기관1 (오투오)	- 출판 콘텐츠 핵심 메타데이터 추출 가능한 자연어 처리 기술 연구 개발 - 생성형 콘텐츠 탐지 기술 및 저작권 분석 기술 - 저작권 분석 및 표절 검출을 위한 한국어 sLLM 기반 자연어 처리 기술 개발	- 카멜 챗봇 - 자연어 생성 AI 기반 기업형 엔터프라이즈 솔루션 납품 (2023년): 스카이라이프, 한국전력공사, 부성형외과 등 공급납품 - 상용서비스 개발 및 제품화 : 메타데이터 분석 및 추천 솔루션 개발 (2023년) - 연구개발과제 수행 능력 : NIA, AI 학습데이터 사업 우수기관 선정 (2022년)
공동연구개발기관2 (고려대 산학협력단)	- 출판 콘텐츠 특화된 한국어 sLLM 설계 및 연구 개발 - sLLM 기반 장르별 창작 AI 기술 연구 개발	- KULLM - 한국어 언어모델 파인튜닝 및 연구개발 기술력 보유 - KoCommonGen v2 - 한국어 기반 언어모델의 일상 지식 능력 평가 벤치마크 데이터셋 구축 관련 연구개발 기술력 보유
공동연구개발기관3 (바이칼레이아이)	- 출판 콘텐츠 및 저작권 Free 콘텐츠 생성을 위한 DB 구축 기술 - 기존 콘텐츠를 활용한 출판 콘텐츠 저작 기술 - 신규 콘텐츠 및 기존 출판 콘텐츠 제공/공유 기술	- NLP / 음성분석 AI 기술에 특화 - 바른 형태소분석기, 바른 개체명분석기 - 유창성센서, 말소리건강점수, 낭독점수, 말명료도센서 등 - 상용서비스 개발 및 제품화 - 자서전출판, 교보생명 다낭, 면접약, 맑은내친구, 맑은동네 - 연구개발과제 수행 능력 - NIA, AI학습데이터 사업 2년연속 우수기관 선정(2건)

3-3. 연구개발 추진일정

1단계-1년차

과제	연구개발내용	1단계(1년차) 월별 추진일정											
		1	2	3	4	5	6	7	8	9	10	11	12
1	장르 선정 및 초기데이터 확보							●	●				
1	sLLM 설계							●	●	●			
2	콘텐츠 구성요소 정의와 분석									●			
2	콘텐츠 구성요소 추출 인스트럭션 데이터셋 설계										●	●	
2	콘텐츠 구성요소 메타 정보 생성 기술 분석											●	●
3	장르별 학습데이터 저작권 확보와 구축								●	●			
4	메타정보, 엔터티, 요소 간 저작 가능형 UI 개발									●	●	●	
5	공유서비스 모델을 위한 아키텍처 설계										●	●	●
5	장르별 세부 속성 정보 DB 구축(저작권,침해요소)										●	●	●

1단계-2년차

과제	연구개발내용	1단계(2년차) 월별 추진일정											
		1	2	3	4	5	6	7	8	9	10	11	12
1	sLLM 2개 개발 및 검증	●	●	●	●	●	●						
2	콘텐츠구성요소 추출 인스트럭션 데이터셋 구축 및 검증		●	●	●								
2	스토리 요약/분석 기술							●	●				
2	표절 검출 기술 개발									●	●	●	●

3	메타정보 기반 분석 DB 구축	●	●	●	●									
3	2차창작용 원천자료 엔티티화(도서,아티클, 문단)	●	●	●	●	●	●							
3	재창작 지원 엔티티 DB 구축	●	●	●	●	●	●	●	●					
3	메타정보 활용 가능 DB 구축	●	●	●										
3	콘텐츠 생성 지원 DB 통합								●	●				
4	2차창작을 위한 요소 추천 알고리즘 개발				●	●	●	●	●					
4	메타 정보 기반 교정·교열 기술 개발					●	●	●	●	●				
5	콘텐츠 공유 서비스 시스템 개발			●	●	●	●	●	●	●	●	●	●	●

2단계-1년차

과제	연구개발내용	2단계 월별 추진일정											
		1	2	3	4	5	6	7	8	9	10	11	12
1	저작권 침해 배제된 장르별 창작기술개발	●	●	●	●	●							
1	스토리 요약/분석 기술 고도화				●	●	●	●	●				
2	요소 파악 기술 개발	●	●	●									
2	표절 검출 기술 고도화			●	●	●							
3	저작권 침해요소별 메타데이터 분석 사례 리포트		●	●	●	●							
3	침해요소 케이스 아카이빙			●	●	●	●						
3	침해요소 DB 공유 기술 개발			●	●	●	●						
4	2차창작을 위한 요소 추천 알고리즘 고도화			●	●	●	●						
4	저작 기술 개발 및 고도화 통합				●	●	●						
5	저작권 침해 정보 케이스 분석형 백엔드 DB 구축 고도화						●	●	●	●	●	●	
5	공유서비스 시스템 고도화								●	●	●	●	●

4. 마일스톤 체계 및 수행계획

1단계

개발 목표	번호	마일스톤 명	목표 일정	수행 기관	주요 가시적 결과물	점검기준	점검방법
1. 출판 LLM	1.1	장르 선정 및 초기데이터 확보	24.07~24.10	컴복스	문서, 데이터 샘플	장르가 적정한가? 수량은 충분한가?	결측치 처리, 데이터 시각화, 편향성 분석
	1.2	sLLM 설계	24.07~24.09	고려대	시스템설계서	모델 설계가 AI 임무에 부합하는가?	리뷰, 프로토타이핑
	1.3	sLLM 2개 개발	25.01~25.05	고려대	sLLM AI 모델 2식	모델의 목표성능 부합하는가? 모델의 방법론은 적합한가?	MLOps, 자체시험도구
	1.3	sLLM 검증	25.04~25.06	고려대	시험보고서	유효성 수준에 부합하는가? 과적합 및 과소적합 여부	성능 평가, 효율성 평가
2. 탐지&분석	2.1	콘텐츠 구성요소 정의와 분석	24.08~24.11	컴복스	요구사항명세서	요구사항 명세 건수 20건 이상 요구사항 명세 상세화 여부	리뷰, 검증회의

개발 목표	번호	마일스톤 명	목표 일정	수행 기관	주요 가시적 결과물	점검기준	점검방법
기술	2.2	컨텐츠 구성요소 추출 인스트럭션 데이터셋 설계	24.10~ 24.11	오투오	시스템설계서	인스트럭션셋 샘플 구성이 충분한가? 인스트럭션셋 설계서 누락 여부	리뷰, 프로토타이핑
	2.3	컨텐츠 구성요소 메타 정보 생성 기술 분석	24.11~ 24.12	오투오	요구사항명세서	요구사항 명세 건수 20건 이상 요구사항 명세 상세화 여부	리뷰, 검증회의
	2.4	컨텐츠구성요소 추출 인스트럭션 데이터셋 구축	25.02~ 25.03	오투오	DB 1식	데이터 구축 건수 1000건 이상 데이터 구축 내용 의미정확성 95% 이상	의미정확성 점검
	2.5	컨텐츠구성요소 추출 인스트럭션 데이터셋 검증	25.04~ 25.04	컴복스	시험보고서	유효성 수준에 부합하는가? 과적합 및 과소적합 여부	성능 평가, 효율성 평가
	2.6	스토리 요약/분석 기술	25.07~ 25.08	오투오	AI 모델 1식	성능지표 수준에 부합하는가? 과적합 및 과소적합 여부	MLOps, 자체시험도구
	2.7	표절 검출 기술 개발	25.09~ 25.12	오투오	AI 모델 1식	성능지표 수준에 부합하는가? 과적합 및 과소적합 여부	MLOps, 자체시험도구
3. 콘텐츠 생성 지원 DB 구축	3.1	콘텐츠 생성 지원 DB 설계	24.09~ 24.12	컴복스	시스템설계서	설계서에 구성요소 충분한가? 설계는 타당한가?	리뷰, 프로토타이핑
	3.2	장르별 학습데이터 저작권 확보와 구축	24.10~ 25.10	컴복스	DB구축계획서	DB 구축 방안이 적절한가? DB 수량은 적절한가?	설계검토, 실현가능성 평가
	3.3	메타정보 기반 분석 DB 구축	25.01~ 25.04	바이칼	DB 1식	DB 응답 성능은 적절한가? DB 데이터는 무결한가?	무결성점검, 가용성점검, 성능 점검
	3.4	2차창작용 원천자료 엔티티화(도서,아티클)	25.04~ 25.10	컴복스	DB 구축 계획서	DB 구축 방안이 적절한가? DB 수량은 적절한가?	설계검토, 실현가능성 평가
	3.5	재창작 지원 엔티티 DB 구축	25.01~ 25.05	바이칼	DB 1식	DB 응답 성능은 적절한가? DB 데이터는 무결한가?	무결성점검, 가용성점검, 성능 점검
	3.6	메타정보 활용 가능 DB 구축	25.01~ 25.03	바이칼	DB 1식	DB 응답 성능은 적절한가? DB 데이터는 무결한가?	무결성점검, 가용성점검, 성능 점검
	3.7	콘텐츠 생성 지원 DB 통합	25.08~ 25.09	바이칼	통합시험결과서	통합시험결과 성공율 95%	테스트케이스검토, 요구사항 추적검사
4. 저작기 술	4.1	저작 기술 요구사항 명세	25.09~ 25.09	바이칼	요구사항명세서	요구사항 명세 건수 20건 이상 요구사항 명세 상세화 여부	리뷰, 검증회의
	4.2	저작 기술 설계	25.10~ 25.10	바이칼	시스템설계서	설계서에 구성요소 충분한가? 설계는 타당한가?	리뷰, 프로토타이핑
	4.3	메타정보, 엔터티, 요소 간 저작 가능형 UI 개발	25.10~ 25.11	바이칼	UI 1식	단위 기능이 정상 동작하는가? 통합 기능이 정상 동작하는가?	단위테스트, 통합테스트
	4.4	2차창작을 위한 요소 추천 알고리즘 개발(1차)	26.04~ 26.05	바이칼	AI 모델 1식(v1)	유효성 수준에 부합하는가? 과적합 및 과소적합 여부	MLOps, 자체시험도구
	4.5	2차창작을 위한 요소 추천 알고리즘 개발(2차)	26.06~ 26.06	바이칼	AI 모델 1식(v2)	유효성 수준에 부합하는가? 과적합 및 과소적합 여부	MLOps, 자체시험도구
	4.6	2차창작을 위한 요소 추천 알고리즘 검증	26.07~ 26.08	바이칼	AI 모델 성능시험보고서	추천 모델의 성능 점검	결과검증, 시험절차검토
	4.7	메타 정보 기반 교정·교열 기술 개발	26.05~ 26.06	바이칼	AI 모델 1식(v1)	유효성 수준에 부합하는가? 과적합 및 과소적합 여부	MLOps, 자체시험도구

개발 목표	번호	마일스톤 명	목표 일정	수행 기관	주요 가시적 결과물	점검기준	점검방법
		(1차)					
	4.8	메타 정보 기반 교정·교열 기술 개발 (2차)	26.07~ 26.07	바이칼	AI 모델 1식(v2)	유효성 수준에 부합하는가? 과적합 및 과소적합 여부	MLOps, 자체시험도구
	4.9	메타 정보 기반 교정·교열 기술 개발 검증	26.08~ 26.08	바이칼	AI 모델 성능시험보고서	교정 교열 기술의 성능 충족이 되는가?	결과검증, 시험절차검토
	4.10	저작기술 통합	26.09~ 26.09	바이칼	통합시험결과서	통합시험결과 성공률 95%	테스트케이스검토, 요구사항 추적검사
5. 콘텐츠 제공 및 공유 기술	5.1	공유서비스 모델을 위한 아키텍처 기획	25.09~ 25.10	바이칼	요구사항명세서	요구사항 명세 건수 20건 이상 요구사항 명세 상세화 여부	리뷰, 검증회의
	5.2	공유서비스 모델을 위한 아키텍처 설계	25.10~ 25.11	바이칼	시스템설계서	설계서에 구성요소 충분한가? 설계는 타당한가?	리뷰, 프로토타이핑
	5.3	장르별 세부 속성 정보 DB 구축 (저작권, 침해요소)	25.11~ 25.12	바이칼	DB 1식	DB 응답 성능은 적절한가? DB 데이터는 무결한가?	무결성점검, 가용성점검, 성능 점검
	5.4	콘텐츠 공유 플랫폼 개발 (1차)	26.03~ 26.05	바이칼	공유플랫폼 1식 (v1)	단위 기능이 정상 동작하는가? 통합 기능이 정상 동작하는가?	단위테스트, 통합테스트
	5.5	콘텐츠 공유 플랫폼 개발 (2차)	26.06~ 26.09	바이칼	공유플랫폼 1식 (v2)	단위 기능이 정상 동작하는가? 통합 기능이 정상 동작하는가?	단위테스트, 통합테스트
	5.6	콘텐츠 공유 플랫폼 통합	26.10~ 26.11	바이칼	통합시험결과서	통합시험 성공률 95%	테스트케이스검토, 요구사항 추적검사

2단계

개발 목표	번호	마일스톤 명	목표 일정	수행 기관	주요 가시적 결과물	점검기준	점검방법
1. 출판 LLM 고도화	1.1	LLM 응용 모델 설계	26.01~ 26.01	고려대	시스템설계서	설계서에 구성요소 충분한가? 설계는 타당한가?	리뷰, 프로토타이핑
	1.2	저작권 침해 배제된 장르별 창작기술개발	26.02~ 26.05	고려대	AI 모델 1식	유효성 수준에 부합하는가? 과적합 및 과소적합 여부	MLOps, 자체시험도구
	1.3	스토리 요약/분석 기술 고도화	26.04~ 26.08	고려대	AI 모델 1식	유효성 수준에 부합하는가? 과적합 및 과소적합 여부	MLOps, 자체시험도구
	1.4	LLM 응용 모델 통합	26.06~ 26.07	고려대	AI 모델 1식	유효성 수준에 부합하는가? 과적합 및 과소적합 여부	MLOps, 자체시험도구
2. 탐지& 분석 고도화	2.1	요소 추출 및 표절 검출 기술 고도화 설계	26.01~ 26.03	오투오	시스템설계서	설계서에 구성요소 충분한가? 설계는 타당한가?	리뷰, 프로토타이핑
	2.2	요소 파악 기술 개발	26.02~ 26.03	오투오	AI 모델 1식	성능지표 수준에 부합하는가? 과적합 및 과소적합 여부	MLOps, 자체시험도구
	2.3	표절 검출 기술 고도화	26.04~ 26.05	오투오	AI 모델 1식	성능지표 수준에 부합하는가? 과적합 및 과소적합 여부	MLOps, 자체시험도구
	2.4	요소 추출 및 표절 검출 기술 고도화 통합	26.05~ 26.05	오투오	AI 모델 성능시험보고서	요소추출 및 표절 검출 모델의 성능 점검	결과검증, 시험절차검토
3. 콘텐츠 생성	3.1	침해요소 분석 설계	26.02~ 26.02	바이칼	시스템설계서	설계서에 구성요소 충분한가? 설계는 타당한가?	리뷰, 프로토타이핑

지원 DB 구축 고도화	3.2	저작권 침해요소별 메타데이터 분석 사례 리포트	26.03~ 26.05	컴복스	리포트 1식	보고서 구성요소 누락 여부	리뷰, 데이터 상호검증
	3.3	침해요소 케이스 아카이빙	26.05~ 26.05	컴복스	DB 1식	DB 응답 성능은 적절한가? DB 데이터는 무결한가?	무결성점검, 가용성점검, 성능 점검
	3.4	침해요소 DB 공유 기술 개발	26.04~ 26.05	바이칼	서비스모듈 1식	단위 기능이 정상 동작하는가? 통합 기능이 정상 동작하는가?	단위테스트, 통합테스트
	3.5	침해요소 통합	26.06~ 26.07	바이칼	DB 1식	DB 응답 성능은 적절한가? DB 데이터는 무결한가?	무결성점검, 가용성점검, 성능 점검
4. 저작기 술 고도화	4.1	저작기술 고도화 설계	26.03~ 26.03	바이칼	시스템설계서	설계서에 구성요소 충분한가? 설계는 타당한가?	리뷰, 프로토타이핑
	4.2	2차창작을 위한 요소 추천 알고리즘 고도화	26.04~ 26.07	바이칼	AI 모델 1식	유효성 수준에 부합하는가? 과적합 및 과소적합 여부	MLOps, 자체시험도구
	4.3	요소 간 관계형 동작 정의 기반저작 기술 개발	26.05~ 26.07	바이칼	AI 모델 1식	유효성 수준에 부합하는가? 과적합 및 과소적합 여부	MLOps, 자체시험도구
	4.4	저작기술 고도화 통합	26.06~ 26.07	바이칼	통합시험결과서	통합시험결과 성공률 95%	테스트케이스검토, 요구사항 추적검사
5. 콘텐츠 제공 및 공유 기술 고도화	5.1	공유플랫폼 고도화 설계	26.04~ 26.05	바이칼	시스템설계서	설계서에 구성요소 충분한가? 설계는 타당한가?	리뷰, 프로토타이핑
	5.2	공유서비스 시스템 고도화	26.04~ 26.08	바이칼	공유플랫폼 고도화 1식 (v1)	단위 기능이 정상 동작하는가? 통합 기능이 정상 동작하는가?	단위테스트, 통합테스트
	5.3	저작권 침해 정보 케이스 분석형 백엔드 DB 구축 고도화	26.08~ 26.09	바이칼	공유플랫폼 고도화 1식 (v2)	공유플랫폼에 대한 단위시험 및 통합테스트 커버리지 70%	단위테스트, 통합테스트
	5.4	공유서비스 고도화 통합	26.10~ 26.11	바이칼	통합시험결과서	통합시험 성공률 95%	테스트케이스검토, 요구사항 추적검사
	5.5	최종 인수시험	26.11~ 26.12	바이칼	시험보고서	인수시험 성공률 95%	성능 평가, 효율성 평가

5. 연구개발성과의 활용방안 및 기대효과

5-1. 연구개발성과의 활용방안

개발성과	활용분야 / 활용방안	설명
저작도구	저자와 일반인을 위한 출판/저작도구로 활용	- 컴복스 및 출판사, 일반 개인들에게 공개 배포. - 특히 컴복스 보유 필진 3700명에게 배포하여 집필 도구로 활용. - 현재 편집자 1인당 1년에 30종을 출간 ➢ 저작도구 활용으로 1인이 100종 이상의 도서를 출간 가능
	실제 출판 과정에 도입	- 컴복스의 주력 시리즈인 ‘이해총서’ 포맷을 실제 집필 과정에 반영 ⇒ 편집 출판 과정 간소화, 비용 절감. - 이 외 다양한 시리즈에 확대, 포맷화 출판 전략에 적용.
공유플랫폼	자산관리시스템을 통한	자산관리시스템을 통해 저작물을 다양한 포맷의 상품으로

개발성과	활용분야 / 활용방안	설명
	출판콘텐츠 2차적 활용 촉진	판매할 수 있으며, 2차적 활용이 용이해짐. 저작권 DBMS로 콘텐츠의 저작권 관리와 수익배분이 용이 → 콘텐츠의 활용성이 높아져 새로운 사업기회로 연결
저작권침해 점검 기술	저작권 제어 기술	- 저작권 침해 위험이 있는 결과물 생성 방지 - 행위 패턴 분석이 필요한 추가적인 점검.
AI 교정/교열 기술	API 판매 또는 기술 이전	- 글쓰기의 장벽으로 저술가가 되지 못한 예비 작가 지원 도구 - 편집 생산성 향상

5-2. 기대효과

분야	기대효과	설명
기술적	한국어 특화 LLM을 통한 수입 대체	- GPT 및 해외 모델의 의존적인 상태임. - 한국어 콘텐츠에 기반한 콘텐츠 생산 능력을 극대화 ➢ 점진적으로 수입 대체 효과 발생
경제산업적	출판업계의 생산성 향상	- 출판업은 대표적인 노동집약적 산업임 - AI기술과 출판기술 접목으로, 자동화의 걸림돌이었던 ‘편집력’을 보완하여 새로운 가치 실현 - 소기업 중심 출판업계(90%가 10인 이하 소기업)의 생산성 향상
	다양한 출판 서비스 개발로 출판업 고도화와 시장 확대, 매출 증대	- 출판 비용 절감으로 더 많은 시도와 출판물 생산 가능해짐. - 이를 활용하여 다양한 출판 서비스와 도서 출간 가능. - 원천 콘텐츠 출판 활성화 → 다양한 콘텐츠가 연쇄적으로 파생
	안정적인 매출을 보장하는 출판BM	판매 상품 다양화, 다각화로 기존 베스트셀러에 의존하는 BM이 아니라 안정적인 매출 수익 구조 운영 가능.
	K콘텐츠 진흥	- 다국어 지원으로 외국인 접속, 콘텐츠 창작과 공유 가능 - 다양한 원천 콘텐츠를 제공하여 K콘텐츠 창작의 산실로 기능. - K콘텐츠 진흥으로 콘텐츠 & 미디어 관련기업의 성장동력으로 작용
문화사회적	창작 활성화 및 창작자 양성	- 다양한 저작도구와 AI저술지원 서비스를 통해 전문 저술가의 생산성 향상과 품질 고도화에 기여. - 일반인의 전문 저술가 진입을 지원하는 등 자아실현과 다양한 직업 선택 기회 제공

5-3. 사업화 전략 및 계획

사업화 모델(BM)	나의 특별한 이야기, mutory.me
사업화 추진 주체	커뮤니케이션박스(주)
시장분석	IT기술과 소량인쇄기술의 발달로 자비출판의 최소 제작부수가 감소하고, 책을 출간하고 싶어하는 저자들을 위한 제작 기회는 증가하였으나, 정식 출판으로 이어져 납본과 서점 판매로 연결되는 경우는 거의 없음. 정식 출간될 수 있는 수준으로 원고의 품질을 높이기

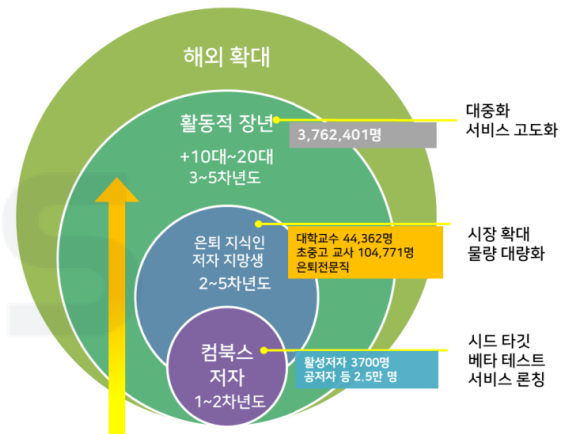
위해서는 상당한 인건비가 투입됨.

- 도서 출간(인건비)과 제작(인쇄, 보관)에 비용이 많이 들어 일반 출판사는 예상 판매량이 높은 도서를 제한적으로 선택하여 출판, 이에 따라 출간 종수가 제한됨.
- K콘텐츠의 인기로 경쟁력 있는 특별한 이야기에 대한 글로벌 요구는 지속적으로 증대.
- 경쟁 서비스 분석

경쟁제품	특징	판매상품
(주)에스프레소북 하루북, 앱스토어 20만 다운	개인/모임 기록을 책으로 묶어 제작 대행. 디자인과 인쇄 대행 서비스에 가까움. → 콘텐츠 2차 활용 불가.	이용자 주문에 따라 인쇄제작비 청구
(주)아이커머 bookpod.co.kr 북팟 웹서비스	자비출판의 온라인 버전. 본문과 표지 디자인 포맷 제공, 소량 인쇄, 유통대행 하지만 상품 가치 없음. → 콘텐츠 2차 활용 불가.	이용자에게 도서 제작비, 편집지원비 청구
스캐터랩 zeta-ai.io '제타'(zeta)	이름, 특징 등을 프롬프트에 입력하면 캐릭터와 나누는 대화 및 스토리에 반영. 이를 통해 로맨스, 판타지, 학원물 등 웹소설이나 웹툰에서 인기있는 다양한 장르를 인터랙티브하게 구현 → 출판물로 상품 가치 없음.	서비스 초기 무료

- 타겟을 점진적으로 확장하여 시장 확대

- 1단계: 출판사 보유 집필진
3700명 타게팅,
mutory.me 홍보,
'나의 이야기' 집필 권유
⇒ 초기에 글감과 매출
안정적 확보
- 2단계: 은퇴 지식인과 저자
지망생, 출판사(BtoB) 홍보
- 3단계: 10대 청소년까지
대상 확장



- 사업 개시 후 정식 출판 종수와 예상 매출

구분	1년차	2년차	3년차	4년차	5년차	합계
출간 종수	50	300	600	900	1200	1200
저자구매(부)	1,000	6,000	6,000	6,000	6,000	25,000
신간판매(부)	2,500	15,000	15,000	15,000	15,000	62,500
서점판매(부)		1,500	9,000	18,000	27,000	55,500
합계(부)	3,500	22,500	30,000	39,000	48,000	143,000
매출(천원)	52,500	337,500	450,000	585,000	720,000	2,145,000

- 2차 창작 예상종수

구분	1년차	2년차	3년차	4년차	5년차	합계
출간 종수	50	300	600	900	1200	1200
보유 아티클	30,000	35,000	43,000	54,000	68,000	68,000
2차 창작 종수	300	18,000	54,000	160,000	480,000	480,000

- mutory.me 서비스 사업화 실현 방안

- 자서전.net 서비스 테스트베드로 활용, '나의 이야기' 수집, 자산관리시스템 가동,

사업화 전략 및
계획

온라인 자동 출판사업 테스트 진행. 1단계(2025년)안에 론칭 운영.

- UI/UX 설계, 디자인 방향 수립에 참고
- 저작도구 프로토타입 테스트
- 교정교열 엔진 프로토타입 테스트
- 이용자 반응 수집 분석
- 사업화 계획 수립에 반영.

○ 2026년 하반기부터 Mutory.me 베타테스트 시작하도록 준비.

● mutory.me 서비스 경쟁력과 전망

- 경쟁력 있는 전문 출판브랜드 없이 책 출간만을 위한 출판사에서 출간되는 책은 정식 출판물과 품질, 유통, 관리 측면에서 차이가 큼. 당사는 ‘나의 이야기’ 전문 출판 브랜드를 육성하여 원고 품질과 상품 수준 관리(포맷화 출판 전략 적용).
- 일반도서 뿐 아니라 전자책, 큰글자책, 오디오북 등 다양한 포맷으로 동시 출간하여 다양한 채널에서 판매, 매출 기회 확대. 당사 보유 무재고 주문출판 시스템을 활용하여 ‘교보문고에서 내 책이 팔리도록’ 일반 도서와 동일한 조건 적용 판매.
- 저자/계약 자산관리시스템을 통해 다양한 2차적 활용 기회 확보 매출 증대.
- 기관 사용자 확보. 콘텐츠 DBMS를 API화하여 BtoB 판매 방안 모색.

끝.

iris