Check for updates
Journal of the Korean Society of Costume
Vol. 76, No. 5 (October 2026) pp. 120-137
Print ISSN 1229-6880
Online ISSN 2287-7827
https://doi.org/10.7233/jksc.2026.76.5.120

조선시대 사료에 근거한 RAG 기반 AI 패션디자인 기획 체계 연구

강규연 · 박선희+
경희대학교 텍스타일·패션디자인학과, 겸임교수 · 이화여자대학교 패션디자인전공, 교수+

RAG-Based AI Fashion Design Planning Framework Grounded in Joseon Dynasty Historical Sources

Gyu Yeon Kang · Sun Hee Park+
Concurrent Professor, Dept. of Textile & Clothing Design, Kyung Hee University
Professor, Dept. of Fashion Design, Ewha Womans University+
(received date: 2026. 7. 14, revised date: 2026. 7. 24, accepted date: 2026. 9. 1)

ABSTRACT

This study examines whether retrieval-augmented generation (RAG) can integrate historical verification into fashion design planning by binding a large language model to public sources. Joseon *Gwanbok* regulations for rank badges and robes are scattered across law codes and royal annals, requiring expert cross-verification for design use. Generative models also remain prone to hallucination, where provenance is essential. Rather than building a new database, the system links three open sources (the Joseon statutes of the National Institute of Korean History, the Institute for the Translation of Korean Classics, and the e-Museum collection) into 1,224 indexed chunks: 133 regulations, 815 records, and 276 artifacts. Under a prespecified protocol, the primary analysis comprised 20 paired design briefs covering 92 fact points, with 80 paired costume-history questions, retrieval evaluation, and cross-model tests supplying supporting evidence. In the briefs, the fact-level hallucination rate was 0.022 and 0.239 with and without RAG, respectively; 83.3% of factual claims with RAG were bound to source identifiers. Factual coverage at the overall fact-point level did not decrease, and the difference was not statistically significant. The direction of hallucination suppression held under GPT-5.4 re-scoring and re-generation, despite variation in absolute rates. Question-level accuracy was 0.838 and 0.438 with and without RAG, respectively. The *Hanja* sub-query linking Korean questions to untranslated classical-Chinese originals was the key retrieval path: Recall@10 was 0.713 in the full configuration and 0.316 without it. The results indicate that source-bound generation can suppress unsupported claims, make design statements traceable, and permit abstention when supporting passages are not retrieved. Designers who are not experts in rank-specific regulations can therefore begin with a traceable brief and verify the cited statutes during planning. These findings establish a prerequisite for source-grounded fashion design; the aesthetic, creative, and practical value of the briefs remains to be evaluated.

Key words: artificial intelligence(인공지능), fashion design(패션디자인), hallucination mitigation(환각 억제), Joseon Dynasty historical sources(조선시대 사료), retrieval-augmented generation(검색증강생성), source traceability(전거 추적 가능성)

Corresponding author: Sun Hee Park, e-mail: zenyul@ewha.ac.kr

I. 서론

1. 연구의 배경 및 필요성

전통복식의 형태와 제도는 문헌과 유물, 도상 사료를 통해 고증되며(Lee, 2005), 이러한 사료는 현대 패션디자인의 원천이 되어 왔다(Park, 2024). 다만 사료를 디자인으로 잇는 작업은 여전히 연구자의 교차검증에 의존하고 있으며, 그 정밀성을 확보하기 위해서는 많은 시간과 노력, 그리고 전문적 판단이 요구된다(K. Lee & Lee, 2026; Lim & Kang, 2026; Park & Cho, 2025). 이러한 부담은 사료를 다루는 연구 안에서보다 오히려 디자인 실무에서 더 뚜렷하게 드러난다. 관복을 디자인의 소재로 활용하려는 디자이너는 한 문양이 어느 품계에 속하는지를 확인하기 위해 『경국대전』에서 『속대전』, 『대전통편』과 『대전회통』으로 이어지는 여러 법전과 실록 기사를 차례로 짚어야 하며, 같은 용어가 사료와 시기에 따라 다른 대상을 가리키는 일이 잦은 탓에 통용되는 설명을 그대로 받아들이면 문양의 귀속 자체가 어긋나기 쉽다. 전통 문양을 영감의 원천으로 삼는 작업에서 이러한 어긋남은 디자인이 내세운 역사적 정당성까지 함께 훼손하므로, 고증의 정확성은 연구의 영역에 그치지 않고 실무에서도 요구되는 조건이 된다.

한편 생성형 대규모 언어모델이 확산되면서 사료의 활용에도 이 새로운 도구가 개입할 여지가 커졌다. 다만 학습된 지식만으로 응답하는 언어모델은 사실과 다른 내용을 그럴듯하게 지어내는 환각(hallucination)을 일으키기 쉬워(Ji et al., 2023), 정확한 전거가 요구되는 사료 영역에 곧바로 적용하기는 어렵다. 이에 응답에 앞서 외부 지식 자원에서 관련 문서를 검색하고 그 근거 위에서 생성하도록 하는 검색증강생성(retrieval-augmented generation, 이하 RAG)이 대안으로 제안되었으며(Lewis et al., 2020), 해외에서는 이를 문화유산과 인문학 분야에 특화하여 적용하는 연구가 이루어지고 있다(Bhatia et al., 2026; Xu et al., 2024; Ye et al., 2025). 그러나 이러한 도구를 한국의 1차 사료, 즉 법전과 실록에 직접 연결하고 그 사실성을 검증한 연구는 저자들이 확인한 범위에서 복식 분야에 아직 보고된 바 없다.

2. 연구 목적 및 연구문제

본 연구의 목적은 검색증강생성을 조선시대 관복의 1차 사료에 직접 결합하여, 공개된 사료 지식 자원을 검색으로 연결한 조건과 검색 근거 없이 학습된 지식만으로 답하는 같은 언어모델의 조건(이하 RAG 미적용)에서 생성된 패션디자인 기획안의 사실성을 비교하고, RAG 조건에서 생성된 사실 주장이 사료 전거에 결속되는 정도를 살피는 데 있다. 본 연구의 관심은 생성형 언어모델을 복식 연구에 쓸 수 있는가라는 일반론이 아니라, 디자이너가 관복 사료를 기획에 활용하는 과정에서 검색의 결합이 생성된 주장의 사실성을 어떻게 바꾸는지와 RAG 조건에서 그 주장의 전거를 어디까지 되짚어 확인할 수 있는지에 있다. 이에 디자인 기획안의 비교를 주 분석으로 두고, 같은 조건의 복식사 질의응답을 짧은 형식에서의 보조 검증으로 수행한다.

본 연구는 패션디자인 기획 과정에 사료 고증을 결합할 수 있는지를 다루되, 검증의 범위는 디자인 기획안이라는 텍스트 산출물의 사실성과 전거 추적 가능성으로 한정한다. 사료의 검색과 생성을 자동으로 잇는 AI 에이전트는 이 과정을 구현한 도구이며, 복식사 질의응답은 짧은 답에서도 사실성의 차이가 나타나는지를 살피는 보조 형식이다. 따라서 본 연구가 확인하는 것은 사료에 근거한 디자인 기획이 성립하기 위한 전제 조건이다.

이를 위한 연구문제는 다음과 같다. 조선시대 관복의 1차 사료를 검색 근거로 결합한 RAG 조건과 같은 언어모델의 RAG 미적용 조건에서 생성된 패션디자인 기획안의 사실성에는 차이가 있는가, 그리고 RAG 조건에서 생성된 사실 주장은 사료 전거에 어느 정도 결속되는가. 이 연구문제를

뒷받침하기 위해 복식사 질의응답에서 정확률과 환각의 양상을 비교하고, 사료 검색이 정답 조항을 실제로 회수하는지와 그 성능이 어느 구성요소에서 비롯하는지를 분리해 살피며, 생성과 채점의 언어모델을 바꾸어도 결론의 방향이 유지되는지를 확인한다. 연구의 방법과 결과 보고, 논의는 모두 디자인 기획안의 사실성과 전거 추적 가능성을 다룬 이 연구문제를 중심으로 구성한다.

3. 연구 범위 및 방법

연구의 범위는 조선시대 관복 제도로 하되, 흉배와 공복, 조복과 제복처럼 품계와 시기에 따라 분화된 규정을 중심으로 삼는다. 이 영역을 연구 대상으로 선정한 것은 서명(書名)과 시기, 전거가 여러 사료에 흩어져 있어 정확한 인용 자체에 전문적인 교차검증이 요구되고, 그만큼 사료 연계의 사실성을 검증하기에 적절하기 때문이다. 자료는 국사편찬위원회의 조선시대법령자료와 한국고전번역원의 한국고전종합DB, 국립중앙박물관의 e뮤지엄이라는 세 공개 지식 자원으로 한정하며, 새 데이터베이스를 구축하는 대신 이미 공개된 자원을 검색으로 연결한다.

방법으로는 먼저 이 세 공개 지식 자원을 검색으로 연결하는 시스템을 구축한다. 이어서 그 검색 결과를 근거로 삼은 산출물의 사실성을 같은 모델의 RAG 미적용 답변과 대응시켜 정량적으로 평가하고, 사료 검색 자체가 정답 조항을 찾아내는 성능도 함께 측정한다. 이때 국역 없이 한문만 전하는 법전 원문을 한국어 질의로 검색할 수 있도록 질의에서 한자어를 추출해 원문에 맞추는 교차언어 검색 경로를 더하고, 그 효과는 이 경로가 없는 구성과 비교하여 확인한다. 아울러 가설과 분석 방법은 결과를 보기 전에 문서로 미리 확정해 두고, 여러 비교를 동시에 수행할 때 요구되는 통계적 보정을 방법 설계에 포함한다. 또한 생성형 언어모델은 종류가 다양하고 빠르게 바뀌는 만큼, 같은 산출물을 계열이 다른 두 모델이 각각 채점하고 생성 모델을 바꾼 재실행을 더하여, 서로 다른 언어모델에서도 검색 결합의 효과가 유지되는지를 함께 확인한다.

II. 이론적 배경

본 장에서는 전통복식 사료를 디자인에 활용해 온 연구와 문화유산·인문학 분야에서 생성형 언어모델 및 검색증강생성(RAG)이 사실성 문제를 다룬 연구를 살펴본다. 이어 조선시대 관복 규범이 여러 사료에 나뉘어 기록된 양상을 검토한다.

1. 전통복식 사료의 디자인 활용과 그 방식

전통복식 연구에서 문헌·유물·도상 사료는 복식의 형태·색채·문양·제도를 고증하는 1차 근거로 기능한다. 조선시대 관복의 제도 변천 역시 법전과 실록, 회화자료의 교차 검토로 규명되어 왔으며(Lee, 2005), 그 성과를 현대 패션디자인으로 잇는 시도 또한 전통과 현대, 서로 다른 문화권의 복식을 대등한 관계에서 결합하는 접근으로 이어졌다(Park, 2024).

사료를 디자인으로 연결하는 실제 작업은 인간 전문가의 교차검증을 근간으로 한다. 대한제국기 대례복의 자수를 재현한 연구는 법령 문헌의 조문별 검토와 유물 실측, 재현 실험을 통합하였고(K. Lee & Lee, 2026), 조선 후기 자색 염색의 재현 연구는 관찬 기록 『상방정례』와 민간 기록 『규합총서』를 비교하여 사료 표현의 모호성까지 재해석하였으며(Park & Cho, 2025), 삿갓을 유형화한 연구는 e뮤지엄의 수록 정보가 소략하여 생긴 분류의 한계를 박물관 현장 실견으로 보완하였다(Lim & Kang, 2026). 이들 연구가 보여 주는 고증의 정교함은 연구자의 많은 시간과 전문적 판단을 요구한다. 본 연구가 주목하는 것은 이 정교한 교차검증을 대체하는 것이 아니라, 그 절차 가운데 사료 검색과 전거 확인의 부담을 디지털 도구와 생성형 도구가 보조할 수 있는가 하는 점이다.

2. 문화유산·인문학 분야의 생성형 AI와 검색증강생성(RAG)

생성형 대규모 언어모델(large language model, 이하 LLM)은 방대한 텍스트로부터 지식을 학습하여 자연어로 응답을 생성하지만, 그 지식은 모델 내부의 수많은 수치 값에 간접적으로 새겨진 것이어서 필요한 사실을 원문 그대로 꺼내 올 수 있는 형태가 아니다. 이로 인해 LLM은 모르는 것을 모른다고 밝히지 못하고(Ji et al., 2023), 답을 지어내는 환각을 일으키는 경향이 있다. 이러한 환각은 사전학습 텍스트의 분포 불균형이나 모델이 사실을 익혀 가는 학습 과정의 역학에서 비롯된 것으로 분석되며(Liu et al., 2026; Zucchet et al., 2025), 사실성이 엄격히 요구되는 사료 영역에서 특히 문제가 된다.

이러한 한계를 보완하는 대표적인 방법이 RAG로, 응답을 학습된 기억이 아니라 검증 가능한 출처에 결속시켜 환각을 줄인다(Lewis et al., 2020). 다만 여러 지식 원천을 결합하면 출처 사이의 불일치로 환각이 오히려 악화될 수 있다는 점도 보고되었다(Wu et al., 2025). 환각을 자동으로 가려내려는 시도도 이어져, 사실성을 시험하는 데이터셋을 자동으로 구성하는 방법(Cao et al., 2026)과 문항 생성부터 평가, 완화까지를 하나로 묶은 도구(Liang et al., 2024)가 제안되었다. 이들은 생성된 진술을 미리 정해 둔 정답 사실과 대조하여 틀린 주장을 세는 방식을 공유하며, 본 연구의 환각률 집계도 같은 계열에 놓인다.

문화유산·인문학 분야에서도 도메인 특화 RAG와 LLM의 적용이 확산되어, 중국 난징의 전통 직조 공예 윈진(雲錦)의 질의응답 시스템이 구축되었고(Xu et al., 2024), 중국 무형문화유산을 위한 특화 언어모델이 개발되었으며(Ye et al., 2025), 이슬람 정전(正典)의 구절에 응답을 결속시키는 질의응답이 시도되었다(Bhatia et al., 2026). 이 흐름의 초점은 문화유산 지식에 대한 접근과 질의응답에 놓여 있다. 아울러 역사 문헌과 고전어는 언어의 특수성이 관건이 되어, 고전 한문에 대한 LLM의 이해도를 측정하는 벤치마크(Y. Wei et al., 2024)와 역사 언어를 위한 자연어처리 도구, 문자 인식의 시대별 오류를 포착하는 평가 틀이 마련되었다(Farsi et al., 2025; Levchenko, 2025). 국내에서도 고전 한문 자료의 전산적 활용의 어려움이 지적되었고(Lee, 2024), 한자 인식 기술의 동향이 정리되었으며(Kim & Choi, 2023), 박물관 소장품 데이터를 생성형 AI로 활용할 때의 신뢰성도 검토되었다(Kim, 2026).

한편 환각 억제라는 측면에서 RAG의 가치는 단답형 질의응답보다 장문 생성에서 더 분명하게 드러난다. LLM은 개방형 질의에 장문으로 응답할 때 사실 오류가 잦아 장문 응답의 사실성을 별도의 벤치마크로 측정하고 자동으로 평가하려는 시도가 이루어졌으며(J. Wei et al., 2024), 이는 복식에 관한 사실이 여럿 누적되는 긴 디자인 기획안에서 사실성 확보가 중요함을 뒷받침한다. 아울러 모르는 것을 확언하는 대신 모른다고 답하는 유보 성향은 모델마다 달라서, 유보가 강한 모델에서는 검색 결합의 이득이 정확률을 끌어올리는 쪽에서, 유보 없이 확언하는 모델에서는 환각을 억제하는 쪽에서 나타날 것으로 예상할 수 있다.

생성형 AI를 디자인에 활용하려는 시도 자체도 활발하다. 패션디자인을 위한 대규모 텍스트-이미지 생성 데이터셋이 구축되었고(Yu et al., 2023), 생성적 디자인 과정에 AI를 방법론적으로 접목하려는 논의가 이어졌으며(Nourian et al., 2024), 인문학의 관점에서 패션 디지털 아카이브를 분석한 연구도 보고되었다(Takahashi et al., 2021). 이 흐름의 초점은 발상과 이미지 생성의 확장에 놓여 있으며, 본 연구가 검토한 범위에서는 국내에서 보고된 생성형 AI 연구 또한 디자인 발상과 교육, 마케팅 활용에 집중되어 있다.

3. 조선시대 관복 제도의 사료와 전거 체계

조선의 관복 규범은 하나의 법전에 집약되어 있

지 않고 여러 사료에 분산·중첩되어 있다. 흉배와 공복의 품계별 규정은 『경국대전』(1485) '예전 의장(儀章)'에 명문화되었고, 『속대전』(1746)에서 문관 당상 흉배가 운학으로 개정되었으며, 『대전통편』(1785)과 『대전회통』(1865)은 새 문양을 제정하지 않고 앞선 규정을 그대로 재수록하였다. 이와 별도로 『국조오례의』 서례(1474)의 '제복도설'과 『조선왕조실록』이 규범의 성립과 변천을 전하며, 흉배 제정을 알리는 단종 2년(1454) 기사가 그 예다. 따라서 한 사실을 정확히 인용하는 데에도 서명과 시기, 전거를 구분하는 전문적인 교차검증이 요구된다.

전거가 여러 층위에 걸쳐 있는 까닭에 용어 역시 문맥에 따라 달리 읽힌다. 예컨대 『국조오례의』 서례 '제복도설'의 품계별 문양은 흉배가 아니라 수(綬)에 관한 것이어서, 수의 운학과 『경국대전』 흉배의 운안, 『속대전』 상복 흉배의 운학이 혼동되기 쉽다. 또한 문관 당상관의 쌍학·당하관의 단학과 같은 정형은 법전 조문이 아니라 관행으로 정립되었으며, 그 시기를 18세기 중후반인 1764년 무렵으로 본 고증이 제시되어 있다(J.-Y. Lee, 2026). 이는 사료 해석의 난점이자, 학습된 지식만으로 응답하는 언어모델이 환각을 일으키기 쉬운 지점이기도 하다.

이러한 모호성을 연구자는 여러 문헌을 교차검증하는 방식으로 해소해 왔다. 『국조오례의』 '빈례'를 중심으로 대일외교의례 복식을 고증한 연구는 대군 기린과 문관 공작·운안·백한, 무관 호표·웅비처럼 당상관에 한정된 『경국대전』 '예전 의장'의 품계별 흉배 규정을 교차 확인하였다(Park et al., 2026). 이러한 인간의 교차검증 절차는 사료 연결형 RAG가 자동화로 보조하고자 하는 대상이며, 동시에 그 결과를 비교해 볼 준거가 된다.

이상의 선행연구를 종합하면, 복식학은 사료의 교차검증으로 고증의 정확성을 확보해 왔으나 전거 확인 절차를 생성형 도구로 보조할 수 있는지는 실증되지 않았다. 문화유산 분야의 RAG 연구는 지식 접근과 질의응답에 집중하여 디자인 산출물의 사실성을 다루지 않았고, 디자인 분야의 생성형 AI 연구는 발상의 확장에 초점을 두어 사료 전거의 사실성을 검증하지 않았다. 이에 본 연구는 조선시대 관복의 1차 사료를 생성 단계에 결속하고, 그 결과로 작성된 디자인 기획안의 사실성을 RAG 미적용 조건과 비교하는 한편, RAG 조건에서 생성된 사실 주장의 전거 결속을 검증한다.

III. 연구방법

본 연구의 시스템은 조선시대 관복 사료에 관한 물음을 받아 법전과 실록, 유물에서 근거를 찾아낸 뒤 그 근거 위에서 디자인 기획안을 작성한다. 이용자의 물음에서 사료 계층의 판정과 검색을 거쳐 기획안 생성에 이르기까지를 자동으로 이어 수행한다는 뜻에서, 본 연구는 이 시스템을 패션디자인 AI 에이전트로 지칭한다. 본 장은 이 시스템의 무엇을 어떤 절차로 확인하였는지를 밝힌다.

1. 연구 설계

본 연구의 평가는 하나의 연구문제를 〈Table 1〉과 같이 주 분석과 세 보조 분석으로 확인하도록 설계하였다. 주 분석은 같은 언어모델에 검색 근거가 주어지는 RAG 조건과 아무 근거 없이 답하는 RAG 미적용 조건을 대응시켜 디자인 기획안의 사실성을 비교하고, RAG 조건에서 생성된 사실 주장의 전거 결속을 확인한다. 복식사 질의응답은 같은 사실성의 차이가 짧은 답에서도 나타나는지를 보조적으로 확인하고, 작동 기제 분석은 생성 단계를 거치지 않는 검색 층위에서 구성요소별 기여를 분리하며, 강건성 분석은 생성과 채점의 언어모델을 바꾸어 주 분석의 결론이 특정 모델의 선택에 종속되는지를 확인한다.

〈Table 1〉 Study Design: Analyses, Conditions, Data, and Metrics

Analysis	Comparison	Data and unit of analysis	Key metrics	Purpose
Primary analysis (design briefs)	With vs. without RAG	20 briefs (92 fact points)	Hallucination rate; source-binding rate	Factuality and traceability of the briefs
Supporting test (costume-history QA)	With vs. without RAG	80 QA items	Question-level accuracy; hallucination rate; source-binding rate	Effect and failure patterns in short answers
Mechanism (retrieval performance)	With vs. without each component	105 retrieval cases	Recall@10·hit@5·nDCG@10	Source of retrieval performance
Robustness (cross-model)	Swapped generator and scorer models	Re-measured per model pair	Gap direction and significance retained	Model independence of the RAG effect

2. 연구 자료 및 대상 선정

분석에 쓰는 자료는 시스템이 검색하는 사료 색인과 평가의 측정 대상이라는 두 층위다. 사료 색인은 〈Table 2〉와 같이 국사편찬위원회 조선시대 법령자료와 한국고전종합DB, e뮤지엄의 세 공개 원천에서 가져온 세 계층으로 구성하였다. 사료 데이터베이스를 새로 구축하는 대신 공개된 지식 자원을 인공지능으로 연결하는 길을 택한 것은, 개별 연구자가 구축의 부담 없이도 공개된 사료를 그대로 재사용할 수 있는지를 함께 확인하기 위해서다.

문헌규정 계층은 대전(大典) 4종, 즉 『경국대전』·『속대전』·『대전통편』·『대전회통』과 예서인 『국조오례의』 서례의 관복 조항을 원문과 국역으로 모아 조항 단위로 나누고 사실마다 서명·편·조와 표기 층위(原·續·增·補)를 전거로 붙인 것으로, 검색 정답의 기준점이 된다. 검색이 걸리는 최소 단위인 문서 조각(chunk)은 여러 품계의 규정이 한 조각에 뒤섞이지도 조항의 문맥이 끊기지도 않을 크기로 잡아야 회수가 정확해지므로, 조의 목차 표제를 품계별 하위 항목으로 세분화하여 이 계층의 색인 조각은 133건이 되었다.

문헌기록 계층은 한국고전종합DB의 실록과 의궤, 문집으로, 법전의 규정이 실제로 어떻게 시행되고 시기에 따라 어떻게 조정되었는지를 확증하는 자리에 해당한다. 관복 핵심 키워드로 검색한 결과에서 중복을 덜어 내어 고전번역서와 고전원문, 실록, 문집총간에서 모두 815개 레코드를 기사 단위로 적재하고, 범위가 넓은 만큼 라이브 조회를 결합하였다. 이 계층에서는 흉배 제정을 전하는 단종 2년(1454)의 실록 기사를 정답의 기준점으로 확정하였다.

끝으로 e뮤지엄에서 관모와 관복, 흉배, 조복, 사모, 단령, 품대에 걸친 관복 유물 276건을 선별해 유물 계층의 색인을 세웠다. 명칭과 설명, 메타데이터에 이미지를 함께 두어 검색으로 유물이 특정되면 법전 조문과 유물 이미지를 한 화면에 나란히 놓고 비교할 수 있게 하였다〈Fig. 1〉. 문헌의 규정과 남아 있는 실물을 함께 살피는 복식 연구의 기본 절차를 화면 구성에 반영한 것이다. 세 계층을 합하면 색인 조각은 1,224건이다.

색인은 국역을 기준으로 삼되 국역이 없는 조항에서는 원문 한자를 그대로 쓰고 한자 보조 질의로 그 빈자리를 메운다. 원문만 전하는 조항은 한국고전종합DB의 국역을 쓰거나 저자들이 접합하여 보완하였는데, 이때 수(綬)의 품계별 문양과 흉배의 품계별 문양은 서로 다른 제도이면서 명칭이 겹치는 대목이 있어 혼동되기 쉬우므로 두 표기가 섞이지 않도록 통제하였다.

〈Table 2〉 Data Layers: Sources, Collection Methods

Layer	Public source	Collection and indexing	Role in this study
Regulations	Joseon statutes database, National Institute of Korean History	Article-level segmentation; source tagging	Reference point for retrieval ground truth
Records	Korean Classics DB, Institute for the Translation of Korean Classics	Local index with live supplementation	Corroborating records; change over time
Artifacts	e-Museum, National Museum of Korea	Local index with live supplementation	Artifact identification; images

측정 대상은 관복 규정의 정답 사실을 모아 둔 사실 원장(ground truth)에서 선정하였다. 원장은 흉배와 공복, 조복과 제복, 관모와 면복, 하위직 복식에 걸친 82개 항목으로, 저자들이 사료 원문을 직접 확인하여 구성하고 각 항목에 서명·조항 전거와 후보 질의를 붙인 것이다. 질의응답에서는 이 82개 항목을 각각 하나의 직답형 문항으로 바꾸고, 원장에 미리 붙여 둔 후보 질의를 그대로 사용하였다. 이 가운데 규정의 부재를 묻는 메타 문항 2개는 제외하고 나머지 80문항을 대응 분석에 사용하였다.

디자인 과제에는 이와 달리 비메타 사실 80개와 규정의 부재를 묻는 메타 항목 2개를 포함한 원장 82개 항목을 사용하였다. 이 항목들이 도메인별로 빠짐없이 포함되도록 브리프(brief) 20편을 구성하되, 같은 항목을 두 편에서 다루는 경우도 허용하였다. 일부 항목이 두 편에 배치되어 브리프별 채점 위치는 92개가 되었고, 이를 사실 지점으로 삼았다.

검색측 기준 케이스 집합은 모두 107케이스로, 검색과 생성이 견뎌야 할 질의 유형을 일곱 차원으로 배분해 구성하였다〈Table 3〉. 비교의 기준선이 되는 직답에 가장 많은 케이스를 두고, 정3품과 종3품이나 서명의 경계처럼 오분류가 일어나기 쉬운 지점, 1485년 규정과 1865년 규정처럼 시기에 따라 조정된 대목, 같은 사실을 다른 말로 묻거나 자연스럽게 물었을 때의 회수, 거짓 전제에 대한 거부와 순환 논리에 대한 저항을 보도록 나머지를 배분하였다. 이 가운데 유물을 정답으로 하는 2건은 문헌 전거 기준의 정답 집합을 적용하기 어려워 경로 판정 정확도만 채점하므로, 정량 채점 대상은 105케이스다.

검색 정답 집합의 관련도는 세 등급으로 설계하였다. 해당 사실을 그대로 담은 조항은 정확 조항, 다른 서명의 재수록 조항이나 문헌기록 계층에서 그 사실을 직접 확증하는 기록은 확증 기록, 같은 법전의 주변 조항은 주변 조항으로 구분하였다. 국역 재수록도 반영하여, 『대전통편』과 『대전회통』의

〈Table 3〉 Composition and Purpose of the Seven-Dimension Retrieval Case Set

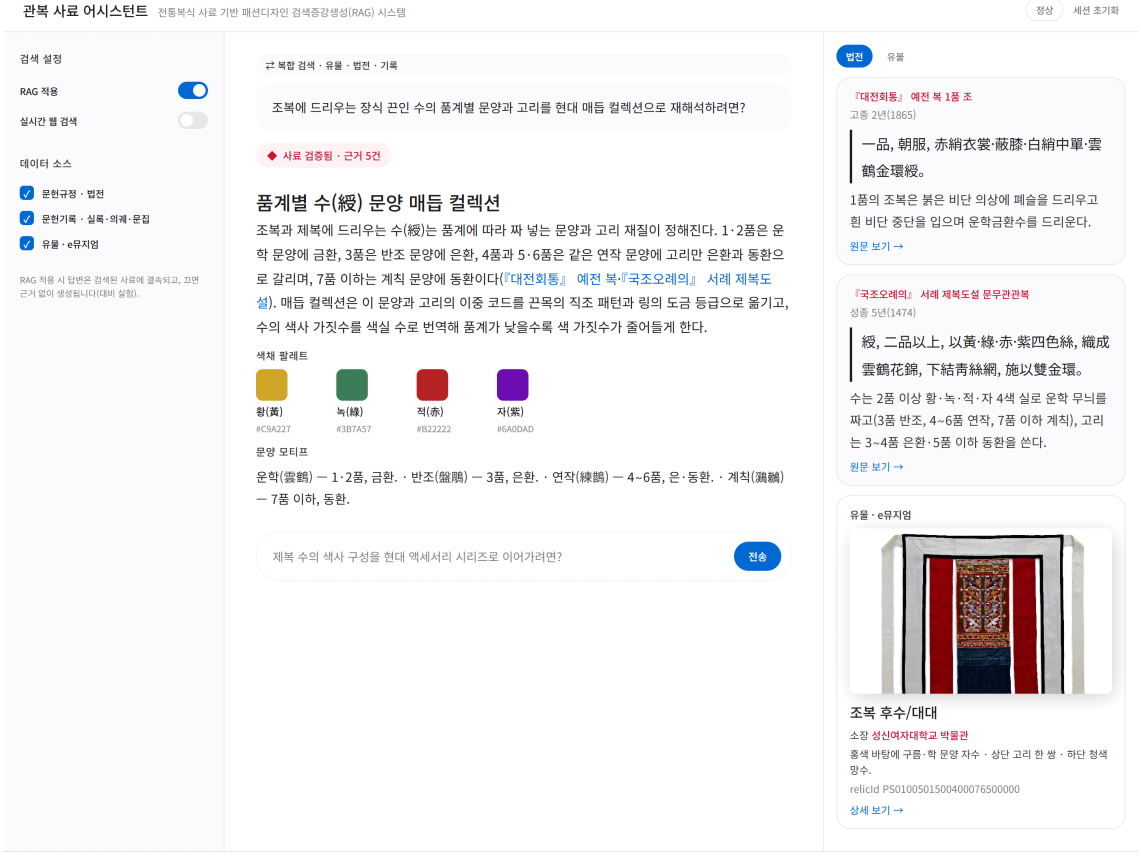
Dimension	Cases	Design purpose
Direct question	33	Baseline
Boundary trigger	6	Misclassification at rank and title boundaries
Change over time	16	Contrast of the same matter across periods
Paraphrase	12	Retrieval for reworded queries
Neutral natural query	13	Retrieval for naturally phrased queries
Trap	21	Rejection of false-premise queries
Circularity	6	Resistance to circular reasoning

국역이 원 조항을 그대로 재수록한 경우에는 이를 정확 조항으로 인정하고 『경국대전』은 전거로 둔다. 정답 집합은 저자들이 사료를 검토하여 직접 선별한 뒤 모든 문서 식별자가 색인에 실재하는지 검증하였다.

3. 비교 조건과 시스템 구성

생성측의 두 조건 가운데 RAG 조건은 검색으로 회수된 사료를 근거로 답하고, RAG 미적용 조건은 아무 근거도 없이 학습된 지식만으로 답한다. 두 조건에서 모델과 질문, 과제, 문항별 반복 횟수는 같게 두었다. 다만 답을 생성하는 방식에는 차이가 있었다. RAG 미적용 조건은 한 번의 실행에서 여러 문항에 연속하여 답하게 한 반면, RAG 조건은 함께 제공하는 사료의 길이를 고려하여 문항마다 따로 답을 생성하였다. 또한 RAG 조건에만 사료에서 근거를 찾지 못하면 답을 지어내지 말고 확인할 수 없다고 밝히라는 지시를 주었다. 따라서 두 조건에서 나타난 차이는 검색 근거 하나만의 효과가 아니라, 사료 제공과 답변 유보를 함께 적용한 RAG 조건 전체의 결과로 해석한다.

〈Fig. 1〉은 이 시스템을 이용하는 화면이다. 물음을 던지면 색채 팔레트와 문양 모티프를 담은 디자인 기획안이 작성되고, 그 옆에 근거로 삼은 사료가 나란히 놓여 문장 하나하나가 어느 조문과



〈Fig. 1〉 User Interface of the RAG System: Statute Article and Artifact Shown Alongside the Design Brief (Sungshin Women's University Museum, n.d.)

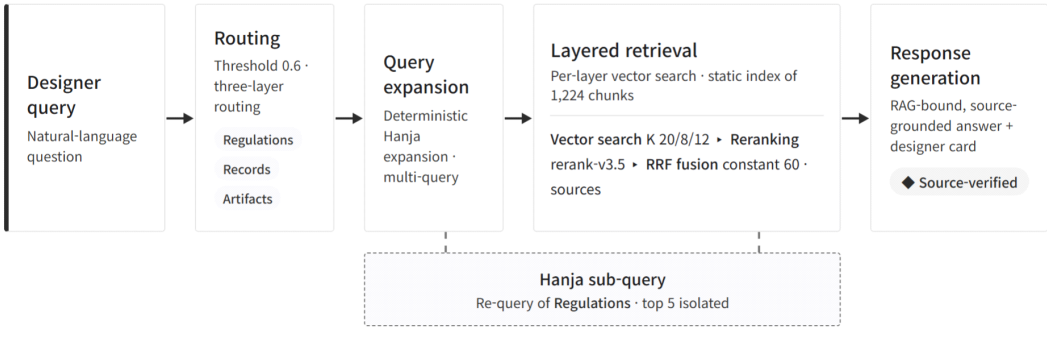
어느 유물에서 나왔는지 따라갈 수 있다. 사료로 확인된 답에는 사료 검증됨 배지가 표시되는데, 여기서 검증은 표기된 사료에서 해당 근거를 찾아 확인하였다는 뜻이며 근거를 찾지 못한 대목은 확인되지 않은 채로 남는다. 화면의 기획안과 사료는 실제 검색 결과로, 법전 근거는 『대전회통』 '예전 복'의 품계별 수(綏) 조항이다. 답을 내놓는 일과 그 답의 출처를 대는 일을 분리하지 않도록 시스템을 구성한 것이다.

시스템은 경로 판정(route), 질의 확장(query expansion), 계층 검색(retrieve), 응답 생성(generate)의 네 단계를 차례로 거치며, 이 주 흐름은 〈Fig. 2〉에 실선으로 나타나 있다. 경로 판정은 질의가 세 계층 가운데 어느 사료를 향한 질문인지를 언어모델의 분류로, 분류가 서지 않으면 키워드 규칙으로 가리되, 신뢰도가 일정 수준에 못 미치면 한 계층을 고르지 않고 모두 조회하여 오분류로 검색이 특정 계층에 갇히는 일을 피한다. 질의 확장은 하나의 질문을 여러 갈래의 검색어로 넓히는 단계로, 관복 용어의 대응 한자를 도메인 사전에서 찾아 붙이는 한자 확장과 언어모델이 같은 질문을 다른 표현으로 바꾸는 다중질의 생성으로 이루어진다.

계층 검색에서는 판정된 계층을 차례로 벡터 검색한다. 이는 문서와 질의의 뜻을 숫자열로 바꾸어 그 가까움으로 문서를 찾는 의미 기반 검색이어서, 글자가 정확히 일치하지 않아도 회수가 가능하다. 1차로 회수된 계층별 상위 문서는 질문과 들어맞는 정도로 다시 채점하여 순서를 바로잡는 재순위(rerank)를 거치고, 서로 다른 계층과 질의에서 나온 여러 순위 목록은 각 목록에서 받은 등수의 역수를 합산하여 여러 목록에 고루 오른 문서를 위로 올리는 상호 순위 융합(reciprocal rank fusion, 이하 RRF)으로 병합한다. 응답 생성은 이렇게 병합된 사료 근거만을 재료로 응답을 만드는 단계다.

한편 한글 질의와 한자 원문 사이의 간극은 별도의 한자 보조 질의 경로를 두어 보완한다. 『경국대전』의 품계별 조항처럼 국역 없이 한문만 전하는 원문에서는 뜻이 같아도 표기 언어가 달라 의미 기반 검색만으로 질의와 문서를 잇기 어렵기 때문이다. 이에 문서 쪽을 번역해 맞추는 대신 질문 쪽에서 언어를 바꾸는 교차언어 검색(query-side cross-lingual retrieval)을 취하여, 한글 질의에서 한자어를 추출해 순수 한자 보조 질의를 만들고 이를 문헌규정 계층에 재조회하여 주 검색 결과와 RRF로 융합한다. 이 경로가 찾아온 문서가 한글 기준 재순위에서 밀려나 탈락하지 않도록 별도 목록으로 확보해 융합에 넘기며, 분기와 재합류의 경로는 〈Fig. 2〉에 점선으로 표시하였다. 그 순효과는 이 경로를 제외한 구성과 비교하여 따로 측정한다.

측정은 한 시점의 색인과 검색 결과를 그대로



〈Fig. 2〉 Architecture of the Source-Linked RAG Pipeline

두고 수행하였다. 검색측의 경로 판정과 한자 확장은 이번 측정에서 언어모델이 아닌 키워드 규칙과 사전으로 처리하였으나, 임베딩과 재순위 단계의 하위 순서는 실행에 따라 달라질 수 있으므로 같은 질의가 언제나 동일한 문서 순서에 이른다고 가정하지 않았다. 검색에는 문장의 뜻을 숫자열로 바꾸는 임베딩(embedding) 모델과 그 숫자열을 담는 벡터 저장소를 쓰며, 단계별로 회수하고 남기는 건수와 융합·판정에 쓴 값은 〈Fig. 2〉에 표시한 대로 측정 전체에서 같게 두었다.

4. 평가 및 분석 방법

평가는 생성측과 검색측을 분리하여 수행하며, 각 지표와 분석이 무엇을 확인하는지는 앞의 연구 설계〈Table 1〉에 대응한다.

1) 생성측 평가

질의응답은 RAG 미적용 조건을 검색과 웹 조회가 금지된 폐쇄형(closed-book)으로 두고 RAG 조건에는 완전 구성의 검색이 회수한 사료를 근거로 주어, 두 조건 모두 두 번씩 실행하되 어느 조건에도 정답은 제시하지 않았다. RAG 조건의 답에는 근거로 삼은 사료의 표기를 요구하여, 그 표기가 실제로 주장을 담은 사료를 가리키는지도 따로 점검한다. 따라서 전거 결속률은 사료 표기를 요구한 RAG 조건에서 평가하고, RAG 미적용 조건은 이 지표의 분석 대상에서 제외하였다.

채점은 두 조건의 답안을 섞어 무작위 순서로 정답과 대조하여 판정한다. 답안에 표기된 사료 번호가 조건을 드러내므로 완전한 맹검 채점은 아니다. 정답은 두 실행이 모두 맞은 문항으로만 세었고, 환각은 어느 한 실행이라도 틀린 값을 확언한 경우로 보았다. 이때 RAG 조건이 근거를 찾지 못해 유보한 답은 정답으로 세지 않으므로, 검색의 실패는 그대로 정확률의 손실이 된다. 한편 RAG 미적용 시행의 정답 여부는 뒤에서 다루는 디자인 과제의 난이도 라벨을 정하는 데도 쓰인다. 따라서 질의응답의 비교는 난이도로 층화하지 않고 전체 문항에서 수행하며, 층화하여 나눈 결과는 탐색적으로만 확인한다.

디자인 과제에서 브리프는 저자가 작성하지만, 난이도 라벨은 저자가 아니라 모델이 정하도록 설계하였다. 위 RAG 미적용 시행에서 생성 모델이 두 번 모두 정답인 사실은 아는 사실(Known)로, 그 밖의 경우는 보수적으로 모르는 사실(Weak)로 판정하며, 이 라벨은 브리프를 확정하기 전에 정해 둔다. 각 브리프는 패션·유니폼·무대의상·액세서리·헤드피스 등의 산출 과제에서 정답 사실 여러 개를 사료 근거로 요구하고 모르는 사실과 아는 사실의 균형을 이루도록 구성하였다. 실제 검색으로 회수된 사료에 결속한 RAG 조건과 RAG 미적용 조건으로 대응시켜 생성하되, 브리프마다 원장의 함정 하나를 심어 환각 유도에 대한 저항을 측정하고 정답은 노출하지 않는다.

디자인 과제의 지표는 모두 사실 원장에 기반한다. 사실 포괄률(coverage)은 대상 사실을 정답과 오답, 부재로 판정한 뒤 정답이 차지하는 비율로 집계한다. 환각률은 두 층위에서 집계하는데, 사실 지점 단위로는 대상 사실을 틀리게 서술한 비율을, 산출물 단위로는 기획안 한 편의 관복 사실 주장 가운데 틀린 주장의 비율을 각각 본다. 전거 결속률은 RAG 조건에서 제시된 주장의 전거가 색인 문서에 실재하고 실제로 그 주장을 담고 있는 비율이다.

생성에는 Claude Opus 4.8을, 채점에는 Claude Sonnet 5를 사용하여 생성과 채점의 모델을 분리하였고, 채점은 어느 모델이든 사실 원장과의 대조로만 이루어지게 하였다. 생성 모델과 같은 계열의 채점자가 그 산출에 후할 수 있다는 자기선호의 우려에 대해서는, 산출 전체를 OpenAI GPT-5.4로 독립 재채점하여 두 채점자의 판정이 겹치는 정도를 원시 일치율과 우연한 일치의 몫을 덜어 낸 Cohen의 κ로 산정하고, 두 채점자가 모두 환각으

로 판정한 지점만 세는 보수적 합의 기준으로 다시 집계하였다. 생성 모델 쪽도 같은 방식으로 교차하여 동일한 문항과 브리프, 동일한 검색 근거를 둔 채 생성만 GPT-5.4로 바꾸어 재실행하였다. 이로써 검색 결합의 효과가 특정 생성 모델에 달린 것인지를 가려, 생성과 채점을 짝지은 네 조합 전부에서 격차가 유지되는지를 확인하였다.

2) 검색측 평가

검색측 채점은 관련도 등급을 점수로 바꾸는 데서 시작하여, 사료 인용에서는 그럴듯한 이웃 조항 여러 건보다 정확한 조항 한 건이 중요하므로 정확 조항·확증 기록·주변 조항을 각각 7점·3점·1점으로 환산하였다. 주지표는 세 가지로, Recall@10은 확증 기록 이상의 정답 문서 가운데 상위 10건 안에 회수된 것의 비율로 정답을 놓치지 않았는지를 보고, hit@5는 정확 조항이 상위 5건 안에 든 케이스의 비율로 가장 확실한 조항을 위에 올렸는지를 보며, nDCG@10은 관련도 높은 문서가 목록의 위쪽에 배치될수록 높아지는 순위 품질 지표다. 이에 더해 실패가 어느 단계에서 비롯했는지를 가리기 위해 경로 판정 정확도와 질의 확장 충실도를 중간 지표로 함께 산출한다.

이와 함께 구성요소의 기여는 완성된 시스템과 해당 요소만 제외한 구성을 동일한 케이스에서 비교하여 분리하며, 관심의 초점은 한자 보조 질의의 순효과에 둔다. 검색을 전면 제거한 조건은 앞서 기술한 생성측 비교에서 다룬다.

3) 통계 처리

통계 처리는 의사반복(pseudo-replication), 즉 같은 케이스에서 나온 측정값을 서로 독립인 표본처럼 세어 표본 수를 부풀리는 문제를 해소하도록 집계 단위부터 설계하였다(Dror et al., 2018). 모든 검정의 1차 단위는 검색측에서 케이스, 생성측에서 질의응답은 대응시킨 문항, 디자인 기획안은 대응시킨 브리프로 잡고, 사실 지점 단위 지표는 브리프 단위로 묶어 비독립성을 보정하였다. 반복 측정은 표본 수를 부풀리는 데 쓰지 않고 시스템 내부의 일관성을 재는 데에만 사용한다.

유의성 검정은 대응 구조와 지표의 성격에 맞추어 선택하였다. 성공과 실패로 나뉘는 이진 지표에는 두 조건 가운데 한쪽에서만 성공한 케이스가 어느 쪽에 더 많은지를 살피는 McNemar 검정을 적용하였다. 비율처럼 값이 연속적으로 변하는 지표에는 같은 표본을 되풀이하여 추출함으로써 조건 차이의 분포를 직접 얻는 대응표본 부트스트랩을, 생성측의 브리프 단위 지표에는 두 조건 사이 차이의 방향과 크기 순위를 함께 보는 Wilcoxon 부호순위 검정을 사용하였다. 두 조건의 비율 차이는 하나의 값으로만 보고하지 않고, 대응 구조를 반영하여 그 차이가 놓일 범위를 나타내는 95% 신뢰구간과 함께 제시하였다(Newcombe, 1998).

p값은 두 조건 사이에 실제 차이가 없다는 가정 아래, 관찰된 차이와 같거나 더 큰 차이가 나타날 확률로, 본 연구는 .05 미만을 통계적으로 유의한 차이의 기준으로 삼았다. 여러 가설을 함께 검정하면 우연히 유의한 결과가 포함될 가능성이 커지므로, 확정 가설에는 작은 p값부터 판정 기준을 조정하는 Holm(1979)의 방법을 적용하고 보정 뒤의 p값을 .05와 비교하였으며, 나머지는 탐색적 분석으로 구분하였다. 보정 범위는 검색측 다섯 구성, 질의응답 두 가설, 디자인 과제 세 가설이며, 방향을 미리 정한 가설은 단측 검정을 썼다. 케이스 수와 차원별 배분은 실제 차이를 찾아낼 확률인 검정력을 산정하여 정하였다(Card et al., 2020). 유의하지 않다는 것은 차이가 없다는 뜻이 아니라 이 자료에서 차이를 확인하지 못했다는 뜻이다.

IV. 연구 결과 및 논의

본 장에서는 앞서 제시한 평가 설계에 따라 측정 결과를 보고하고 그 의미를 논의한다. 모든 수치는 본 연구에서 수행한 단일 측정 계열에서 얻

은 것이다. 먼저 연구문제가 직접 묻는 디자인 기획안의 사실성과 전거 추적 가능성을 살펴보고, 복식사 질의응답과 검색 성능을 보조 분석으로 이어서 확인한다.

1. 생성 산출물의 사실성과 전거 추적

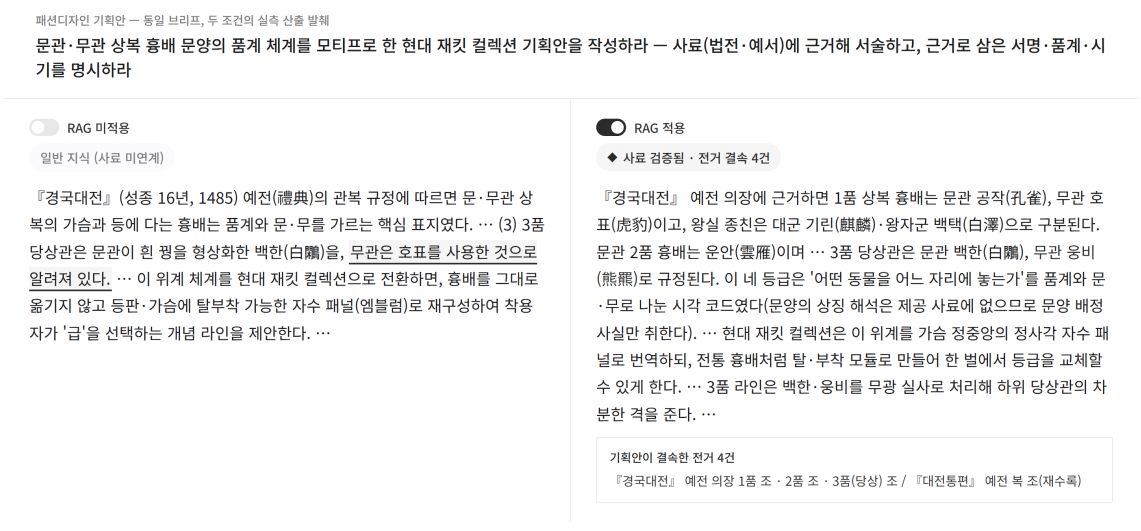
앞서 정의한 RAG 조건과 RAG 미적용 조건에 같은 과제를 주어 생성한 디자인 기획안을 대응시켜 비교하였다. 〈Fig. 3〉은 정량 측정 과정에서 실제로 생성된 두 기획안을 나란히 보여 준다.

1) 패션디자인 기획안

디자인 기획안 한 편에는 문양과 품계, 재질과 시기, 서명에 관한 사실 주장이 여럿 담기므로, 사실 하나를 묻는 짧은 답보다 잘못된 귀속과 혼동이 디자인 판단에 누적될 가능성이 크다. 실제로 RAG 미적용 조건에서는 별감과 나장 같은 하위직 복식에 존재하지 않는 명칭을 만들거나 품계에 따른 흉배 문양을 잘못 귀속하는 사례가 나타났다. 〈Fig. 3〉의 현대 재킷 컬렉션 기획안에서도 RAG 미적용 조건은 무관 3품 당상의 웅비(熊羆)를 인접 품계의 호표(虎豹)로 바꾸어 서술한 반면, RAG 조건은 『경국대전』의 웅비 규정을 제시하고 오른쪽 패널에 사료 검증 상태와 네 건의 조문 전거를 함께 표시하였다. 이러한 오류가 그대로 모티프 선정에 반영되면 기획안이 내세운 역사적 정당성까지 훼손될 수 있으므로, 품계와 문양의 귀속을 사료에서 되짚어 확인할 수 있다는 점이 디자인 기획 과정에서 나타난 뚜렷한 차이이다.

두 조건의 산출물을 사실 원장과 대조한 결과는 〈Table 4〉와 같다. 전체 92개 사실 지점을 기준으로 한 환각률은 RAG 조건 0.022, RAG 미적용 조건 0.239로 나타나 검색 근거가 결합된 기획안에서 약 10분의 1 수준으로 낮아졌다. 기획안 한 편이 내놓은 사실 주장 가운데 틀린 주장이 차지하는 비율의 20편 평균값인 편당 환각률도 각각 0.037과 0.256이었으며, 브리프 20편을 대응시킨 차이 +0.220은 단측 p=.001로 유의하였다. 결과표의 값은 모두 소수점 넷째 자리에서 반올림한 것이다.

전거 결속률은 근거로 삼은 사료를 표기하도록 한 RAG 조건에서 평가하였다. 이 조건에서는 기



〈Fig. 3〉 Comparison of Fashion Design Briefs Without RAG (Left) and With RAG (Right): Correcting the Rank-Badge Attribution

〈Table 4〉 Factuality and Source Traceability of Fashion Design Briefs

Generator	Scorer	Metric	With RAG	Without RAG
Opus 4.8	Sonnet 5	Fact-point hallucination rate	0.022	0.239
		Per-brief hallucination rate	0.037	0.256
		Factual coverage	0.826	0.717
		Source-binding rate	0.833	–
	GPT-5.4	Fact-point hallucination rate	0.163	0.348
		Per-brief hallucination rate	0.093	0.237
		Factual coverage	0.685	0.598
		Source-binding rate	0.801	–
GPT-5.4	Sonnet 5	Fact-point hallucination rate	0.207	0.598
		Per-brief hallucination rate	0.167	0.449
		Factual coverage	0.663	0.337
		Source-binding rate	0.870	–

획안이 제시한 사실 주장 90건 가운데 75건, 즉 83.3%가 법전 조문 식별자 수준의 전거에 결속되었다. RAG 미적용 조건에는 사료 전거의 표기를 요구하지 않았으므로 이 지표의 분석 대상에서 제외하였으며, 〈Table 4〉의 '–'는 이에 해당한다. RAG 조건은 회수된 사료가 없는 사실을 일부 유보하였지만, 전체 사실 지점을 기준으로 한 사실 포괄률은 RAG 조건 0.826, RAG 미적용 조건 0.717로 나타났다. 다만 두 조건의 차이는 p=.062로 유의하지 않았으므로, 포괄률이 더 높게 나타난 방향은 관찰되었으나 이 자료만으로 두 조건의 차이를 확인하지는 못하였다. 채점자와 생성 모델을 바꾼 강건성 분석에서도 환각 억제의 방향은 유지되었다. 계열이 다른 GPT-5.4의 독립 재채점에서 브리프를 대응시킨 환각률의 차이는 +0.144로 p=.010이었으며, 생성 모델을 바꾸거나 생성과 채점을 함께 바꾼 재실행에서도 RAG 미적용 조건의 환각이 더 높았다. 이 가운데 생성과 채점을 모두 바꾼 조합은 표에 싣지 않았고, 표에 실은 두 강건성 조합의 전거 결속률 0.801과 0.870은 각각 주장 171건 가운데 137건과 154건 가운데 134건에 해당한다.

이 결과의 실무적 의의는 사료에 근거한 디자인 기획이 성립하기 위한 전제 조건으로서 생성된 주장의 사실성과 전거 확인 가능성을 함께 검토한 데 있다. 흉배와 품계 규정을 온전히 익히지 못한 디자이너도 기획안을 작성하는 과정에서 각 주장의 조문 근거를 함께 받아 확인할 수 있기 때문이다. 다만 RAG 조건에도 원령포(圓領袍)의 '圓'을 '團'으로 적는 것과 같은 생성 단계의 잔여 혼동이 남았고, 회수된 사료가 없을 때에는 일부 사실을 다루지 못하는 공백이 생겼다. 난이도 층화에서는 널리 알려지지 않은 사실의 환각 차이만 유의하였고 나머지 가설은 예측한 방향에 머물렀다.

2) 복식사 질의응답

짧은 형식에서도 검색 결합에 따른 사실성의 차이가 나타나는지를 확인한 결과는 〈Table 5〉와 같다. 문항 정확률과 문항 환각률은 모두 대응시킨 80문항을 분모로 하였다. 두 번의 실행이 모두 정답이어야 맞은 것으로 세었을 때 문항 정확률은 RAG 조건 0.838, RAG 미적용 조건 0.438이었고, 그 차이는 p<.0001로 유의하였다. 이 형식에서 검색의 이득은 모델이 모르던 세부 규정을 전거와 함께 정답으로 바꾸어 놓는 데서 뚜렷하였다.

환각률은 RAG 조건 0.025, RAG 미적용 조건

〈Table 5〉 Factuality Comparison in Costume-History Question Answering

Generator	Scorer	Metric	With RAG	Without RAG
Opus 4.8	Sonnet 5	Question-level accuracy	0.838	0.438
		Question-level hallucination rate	0.025	0.087
		Source-binding rate	1.000	–
	GPT-5.4	Question-level accuracy	0.825	0.362
		Question-level hallucination rate	0.025	0.087
		Source-binding rate	0.935	–
GPT-5.4	Sonnet 5	Question-level accuracy	0.825	0.150
		Question-level hallucination rate	0.013	0.625
		Source-binding rate	0.924	–

0.087이었으나 그 차이는 p=.090으로 유의하지 않았다. RAG 미적용 조건의 실패가 틀린 값의 확언보다 모른다는 답이나 부분 답으로 나타나 환각으로 셀 사례가 적었기 때문이다. 그러나 생성 모델을 유보 없이 답하는 GPT-5.4로 바꾸면 RAG 미적용 조건의 실패가 확언된 오답으로 바뀌었고, 정확률과 환각률의 차이는 생성과 채점 모델을 함께 바꾼 조합에서도 같은 방향으로 유지되었다. 전거 결속률은 문항이 아니라 답을 단위로 세어, 80문항을 두 번 실행한 160개의 답 가운데 전거 판정이 가능했던 답을 분모로 하였다. RAG 미적용 조건에는 사료 전거의 표기를 요구하지 않았으므로 이 지표의 분석 대상에서 제외하였으며, 〈Table 5〉의 '–'는 이에 해당한다. 원 측정의 RAG 조건에서는 156개의 답이 모두 전거에 결속하여 1.000이었고, 모델을 바꾼 두 조합에서도 153건 가운데 143건과 158건 가운데 146건이 결속하여 0.92를 넘었다. 반대로 검색이 정확 조항을 회수하지 못한 경우에는 RAG 조건이 답을 지어내는 대신 확인 불가로 유보하였고, 일부 답은 값이 맞더라도 법전의 서명을 잘못 귀속하였다. 따라서 질의응답의 보조 검증에서는 RAG 조건의 정답률이 높게 나타났고, 이 조건에서 생성된 답의 전거도 되짚어 확인할 수 있었다. 아울러 회수 실패는 유보와 귀속 착오로 남을 수 있음을 보여 준다.

종합하면 여러 사실이 누적되는 디자인 기획안에서는 검색 근거를 결합한 조건의 환각률이 낮게 나타났고, RAG 조건에서 생성된 사실 주장 가운데 높은 비율이 전거에 결속되었다. 짧은 질의응답에서는 세부 규정을 정답으로 바꾸는 양상이 먼저 드러났다. 생성 모델의 유보 성향에 따라 실패의 형태는 달라졌지만, 네 모델 조합에서 RAG 조건의 환각률은 미적용 조건보다 낮게 나타났다. 『경국대전』 '예전 의장'의 흉배 규정을 교차 확인한 선행연구와 본 연구의 정답 기준은 일치하지만, 사료 표현의 모호성을 재해석하거나 유물을 직접 실견하여 분류를 보완하는 일은 검색 결합으로 대신할 수 없다(Lim & Kang, 2026; Park et al., 2026). 본 연구의 기여는 질의응답을 중심으로 논의되어 온 사료 특화 RAG를 한국의 1차 사료와 패션디자인 기획안이라는 실무 산출의 형식으로 확장하고, 그 사실성과 전거 추적 가능성을 검증한 데 있다.

2. 사료 검색 성능과 한자 보조 질의

검색 성능은 주 분석에서 확인한 사실성이 어떤 회수 경로 위에서 성립하는지를 살피는 보조 분석이다. 회수된 문서 목록을 미리 확정한 정답 집합과 대조하는 채점은 프로그램으로 처리하였으며, 유물을 정답으로 하는 2건을 제외한 105케이스의 결과는 〈Table 6〉과 같다. 표의 Recall@10과 nDCG@10은 케이스별 값의 105케이스 평균값이다. 전체 107

〈Table 6〉 Retrieval Performance with and without the Hanja Sub-Query

Configuration	Recall@10	nDCG@10	hit@5
Full configuration	0.713	0.447	0.676
Without Hanja sub-query	0.316	0.274	0.343

케이스 가운데 거짓 전제를 담은 함정 21건과 순환성·경계 촉발 각 6건을 합한 33건이 약 30%를 차지하므로, 이 결과는 찾기 쉬운 직답형 질의만으로 얻은 것이 아니다.

가장 중요한 구성요소는 한자 보조 질의였다. 이 경로를 제외하면 Recall@10은 0.713에서 0.316으로 낮아져 완전 구성의 절반에 미치지 못하였고, hit@5도 정확 조항이 상위 5건에 든 케이스가 105건 가운데 71건에서 36건으로 줄어 0.676에서 0.343으로 낮아져 절반에 가까운 수준으로 나타났다. 다시 말해 한자 보조 질의를 갖춘 구성의 상위 10건 정답 회수는 이를 제외한 구성의 두 배 이상이었다. 정확 조항의 상위 회수에서 나타난 차이는 $p<.0001$로 유의하였으며, 다섯 구성을 비교한 결과 유의한 차이를 보인 요소도 한자 보조 질의뿐이었다. 이는 국역이 없는 한자 법전 원문을 한국어로 찾는 경로가 디자인 기획안의 사료 근거를 확보하는 데 핵심임을 뜻한다. 다만 조항이 짧은 하위직 규정과 자연어 질의의 회수에는 한계가 있었다.

V. 결론 및 제언

1. 결론

본 연구는 공개된 조선시대 관복 사료를 검색으로 연결하여 패션디자인 기획안에 사료 고증을 통합할 수 있는지를 검토하였다. 주 분석인 디자인 기획안 생성에서는 사실 지점 환각률이 RAG 미적용 조건의 약 10분의 1 수준으로 낮게 나타났고, RAG 조건에서 생성된 사실 주장의 83.3%가 색인에 실재하는 조문 전거에 결속되었다. 복식사 질의응답에서도 문항 정확률이 RAG 조건 0.838, RAG 미적용 조건 0.438로 나타나 같은 방향을 보였다. 이로써 서론에서 세운 연구문제에 대해, 사료 검색을 결합한 조건에서는 디자인 기획안의 환각률이 낮게 나타났고, RAG 조건에서는 생성된 사실 주장의 전거 결속이 확인되었다. 검색 성능에 가장 크게 기여한 경로는 한자 보조 질의였다. 이 경로를 빼면 상위 10건까지의 정답 회수가 0.713에서 0.316으로 낮아졌으며, 한자 원문을 한국어로 찾는 경로가 사료 기반 기획의 중요한 조건임을 보여 주었다. 여기서 환각 억제는 생성된 서술에서 사실 원장과 어긋나는 확언이 크게 줄어들었다는 뜻이다. 전거 결속은 이와 서로 관련되지만 같은 개념은 아니어서, 그 서술이 표기된 사료에 실제로 담겨 있는지를 되짚어 확인할 수 있음을 가리킨다. 따라서 본 연구는 이 결과를 검색의 결합이 역사적 사실성 자체를 보장한다는 뜻이 아니라, 사실 원장에 비추어 환각이 줄어들고 RAG 조건에서 생성된 주장의 전거를 되짚을 수 있게 된 것으로 해석하였다.

전통복식을 현대 디자인으로 잇는 작업에는 기획에 앞서 관련 문헌과 유물을 폭넓게 살피거나, 기획안을 마련한 뒤 모티프의 귀속과 시기를 다시 확인하는 고증이 요구된다. 본 연구의 시스템은 이러한 고증을 대체하지 않고, 사료 검색과 전거 확인을 기획안 작성 과정에 연결하여 생성된 서술의 근거가 되는 조문과 유물을 함께 제시한다. 관복 제도의 세부 전거를 온전히 익히지 못한 연구자와 디자이너도 조문에 결속된 기획안을 바탕으로 모티프와 품계의 귀속을 확인하며 기획을 이어 갈 수 있다. 본 연구에서 관찰된 환각 억제와 전거 결속, 근거가 회수되지 않을 때의 유보가 이러한 작업 방식의 출발 조건이 갖추어질 수 있음을 시사한다는 데 실무적 의의가 있다.

여러 사실이 누적되는 긴 기획안에서 환각 억제의 효과가 더 뚜렷하였다는 결과는 장문 산출이 필요한 디자인 실무에서 사실성 확인이 중요함을 보여 준다. 생성 모델의 유보 성향에 따라 짧은 답에서 실패가 드러나는 방식은 달랐으나, 검색 근거를 결합한 조건에서 환각률이 낮게 나타나는 방향은 생성과 채점 모델을 바꾼 조합에서도 유지되었고 전체 사실 지점을 기준으로 한 사실 포괄률도 RAG 미적용 조건보다 낮지 않았다. 법전과 실록이라는 한국의 1차 사료를 직접 연결하고 디자인 기획안의 사실성을 정량적으로 검증한 연구는 저자들이 확인한 범위에서 복식 분야에 보고된 바 없었다. 본 연구는 생성형 도구의 신뢰성을 환각 억제와 전거 결속이라는 측정 가능한 속성으로 옮겨, 복식학에서 사료 기반 패션디자인 기획을 검증 가능하게 논의할 방법적 기반을 마련하였다는 데 학술적 의의가 있다.

2. 연구의 제한점

채점자와 생성 모델을 바꾼 교차에서도 전거 결속률은 질의응답에서 0.92에서 1.00, 기획안 생성에서 0.801에서 0.870에 이르는 값으로 나타났다. 이는 검색 근거를 결합한 생성에서 전거 결속이 모델 조합에 걸쳐 유지되었음을 보여 주었다. 한편 두 비교 조건의 답변 생성 방식에는 차이가 있었다. RAG 미적용 조건은 한 번에 여러 문항에 연속하여 답하였지만, RAG 조건은 함께 제공하는 사료의 길이를 고려하여 문항마다 따로 답을 생성하였다. 또한 사료에서 근거를 찾지 못하면 답을 유보하라는 지시는 RAG 조건에만 주어졌다. 따라서 두 조건에서 나타난 차이는 검색 근거 하나만의 효과가 아니라, 사료 제공과 답변 유보를 함께 적용한 RAG 조건 전체의 결과로 해석하였다.

본 연구의 평가는 생성된 서술의 사실성과 제시된 전거의 결속 여부를 대상으로 하였으며, 기획안의 심미성이나 창의성, 디자인 실무에서의 유용성은 별도의 전문가 평가를 요하는 문제로 다루지 않았다. 또한 사실 원장과 정답 집합, 관련도 판정은 저자들이 사료를 확인하여 구성한 것으로 복식사 전문가의 독립 검증을 거치지 않았으므로, 본 연구의 환각 억제와 전거 결속은 이 원장과 정답 집합을 기준으로 판정된 결과이다. 검색의 측면에서 남는 과제는 사료의 규모에서 비롯된다. 법전과 실록, 문집으로 이어지는 조선시대 고전 사료는 규모가 방대하고 국역의 층위도 고르지 않아, 한자 보조 질의로 회수를 크게 끌어올린 뒤에도 조항이 짧은 세부 규정을 최상위에 올리거나 자연어 질의의 회수를 고르게 높이는 데에는 보다 정교한 검색 기법이 필요하다. 검색 지표는 미리 정해 둔 정답 집합을 기준으로 채점되므로, 정답으로 등재되지 않은 문서는 실제로 관련이 있더라도 관련 없는 문서로 처리된다. 이 때문에 본 연구는 지표의 절대적인 수치보다 같은 정답 집합 위에서 구성을 바꾸어 얻은 값들을 서로 비교하여 결과를 해석하였다. 아울러 이 접근은 국역과 전거, 공공 이미지 같은 공개 자료 기반 위에서 성립하므로, 다른 시대나 복식 영역으로 넓히려면 그에 상응하는 자료 기반이 갖추어져 있어야 한다.

3. 제언

이러한 제한점은 각각 구체적인 후속 과제로 이어진다. 먼저 결과의 검증에서는 복식사 전문가의 독립 검증을 더하고, 생성과 채점을 짝지은 네 조합의 교차를 더 많은 모델 계열로 확대하면 결과의 일반화를 한층 두텁게 뒷받침할 수 있다. 평가의 범위도 넓힐 여지가 있어, 생성된 서술의 사실성과 제시된 전거의 결속 여부를 다룬 본 연구의 평가를 기획안의 심미성과 창의성, 실무 유용성에 대한 디자이너·연구자 평가로 확장하는 일이 다음 단계의 과제다. 한편 검색에서는 하위직 조항의 한자 대응을 보완하고 자연어 질의를 위한 기법을 강화하는 동시에, 결정적 경로로 측정한 계층 경로 판정에 언어모델 분류를 결합하였을 때의 효과를 확인함으로써 회수를 더 높일 수 있다. 유물 검색

은 문양이나 유물 사진을 질의로 받는 이미지 검색으로 확장하여 시각 임베딩 색인을 병용하면 설명문 밖 문양의 세부까지 대조할 수 있으며, 나아가 본 연구가 연결한 사료 자원을 다른 시대와 복식 영역으로 넓히면 공개 지식 자원을 활용한 사료 연계 생성의 적용 범위를 확장할 수 있을 것이다. 본 연구의 시스템 구성과 평가 설계는 전통복식 사료를 활용한 생성형 도구를 검토하려는 후속 연구와 사료에 근거한 디자인 기획 실무에 기초자료로 활용될 수 있을 것으로 기대된다.

References

Bhatia, G., Mubarak, H., Jarrar, M., Mikros, G., Zaraket, F., Alhirthani, M., ... Alam, F. (2026). From RAG to agentic RAG for faithful Islamic question answering. In *Findings of the Association for Computational Linguistics: ACL 2026* (pp. 26469-26488). Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.findings-acl.1317

Cao, Z., Yang, Y., Li, X., & Zhao, H. (2026). AutoHall: Automated factuality hallucination dataset generation for large language models. *IEEE Transactions on Audio, Speech and Language Processing, 34*, 184-195. https://doi.org/10.1109/TASLPRO.2025.3635038

Card, D., Henderson, P., Khandelwal, U., Jia, R., Mahowald, K., & Jurafsky, D. (2020). With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 9263-9274). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.745

Dror, R., Baumer, G., Shlomov, S., & Reichart, R. (2018). The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1383-1392). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1128

Farsi, F., Fazel, P., Haghighi, S., Sabouri, S., Goshtasb, F., Hajipour, N., ... Sameti, H. (2025). ParsiPy: NLP toolkit for historical Persian texts in Python. In *Proceedings of the Second Workshop on Ancient Language Processing (ALP 2025)* (pp. 137-149). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.alp-1.17

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65-70. https://www.jstor.org/stable/4615733

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), Article 248. https://doi.org/10.1145/3571730

Kim, N. (2026). A study on the reliability of museum collection data as generative AI training data: An assessment of eMuseum based on the L·Q dual-condition framework [*생성형 AI 학습데이터로서 박물관 소장품 데이터의 신뢰성 연구: 이원적 분석틀(L·Q)을 통한 e뮤지엄 진단*]. *Journal of Museum Studies*, (51), 57-86. https://doi.org/10.22884/joksms.51.202606.003

Kim, W.-J., & Choi, J.-Y. (2023). Current status and prospects of Chinese character recognition models in Korea [*인공지능 기반 한자 인식 기술 현황과 과제*]. *Eomunyeongu, 51*(4), 313-341. https://doi.org/10.15822/skllr.2023.51.4.313

Lee, E.-J. (2005). Development process of Sibok and Sangbok for officer in Joseon Dynasty [*조선시대 백관의 시복과 상복 제도 변천*]. *Journal of the Korean Society of Costume, 55*(6), 38-50. https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART001113438

Lee, J.-Y. (2026). The establishment and evolution of the civil official Hyungbae system in the Joseon Dynasty [*조선시대 문관 흉배제도의 성립과 전개*]. *Journal of Korean Traditional Costume, 29*(1), 17-35. https://doi.org/10.16885/jktc.2026.03.29.1.17

Lee, K., & Lee, J. (2026). A study on embroidery for the reproduction of court dress according to the regulations of the Gungnaebubonbu(宮內府本府) and Yesigwon(禮式院) in the Daehan Empire [*대한제국기 궁내부 본부 및 예식원 예복 제도의 대례복 재현을 위한 자수 연구*]. *Journal of the Korean Society of Costume, 76*(2), 67-81. https://doi.org/10.7233/jksc.2026.76.2.067

Lee, M. (2024). Difficulties in using Korean classical Chinese materials in text mining-based historical research [*텍스트 마이닝 기반 역사 연구에서 한국 고전 한문 자료 활용의 어려움*]. *The Daegu Historical Review, 155*, 1-29. https://doi.org/10.17751/DHR.155.1

Levchenko, M. A. (2025). Evaluating LLMs for historical document OCR: A methodological framework for digital humanities. In *Proceedings of the First Workshop on Natural Language Processing and Language Models for Digital Humanities (LM4DH 2025)* (pp. 75-85). INCOMA. https://doi.org/10.26615/978-954-452-106-6-007

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive

NLP tasks. *Advances in Neural Information Processing Systems, 33,* 9459-9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Liang, M., Arun, A., Wu, Z., Munoz, C., Lutch, J., Kazim, E., ... Treleaven, P. (2024). THaMES: An end-to-end tool for hallucination mitigation and evaluation in large language models. In *NeurIPS 2024 Workshop on Socially Responsible Language Modelling Research (SoLaR).* https://openreview.net/forum?id=GMNDgcjPfy

Lim, J., & Kang, S.-Y. (2026). A study on the typology and formative features of Korean Satgat through artifacts [*유물을 통해 본 한국 삿갓의 유형화와 조형성 연구*]. *Journal of the Korean Society of Costume, 76*(1), 183-202. https://doi.org/10.7233/jksc.2026.76.1.183

Liu, L., Lv, K., Chen, H., Zhang, W., Wang, Y., Liu, S., ... Zheng, B. (2026). PretrainRL: Alleviating factuality hallucination of large language models at the beginning. In *Findings of the Association for Computational Linguistics: ACL 2026* (pp. 18298-18310). Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.findings-acl.910

Newcombe, R. G. (1998). Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine, 17*(22), 2635-2650. https://doi.org/10.1002/(SICI)1097-0258(19981130)17:22<2635::AID-SIM954>3.0.CO;2-C

Nourian, P., Azadi, S., Uijtendaal, R., & Bai, N. (2024). Augmented computational design. In N. Abbasabadi & M. Ashayeri (Eds.), *Artificial intelligence in performance-driven design: Theories, methods, and tools towards sustainability* (pp. 1-30). Wiley. https://doi.org/10.1002/9781394172092.ch1

Park, D., Cho, W., & Cho, K. S. (2026). A study on the costumes used in diplomatic ceremonies for Japanese envoys in early Joseon: Focusing on the Suingukseopyeui in Gukjo-Oryeui [*조선 초기 對日賓禮 복식 연구 -『국조오례의』 受隣國書幣儀를 중심으로 -*]. *Journal of the Korean Society of Costume, 76*(2), 82-102. https://doi.org/10.7233/jksc.2026.76.2.082

Park, H. (2024). An exploratory study on equitable cross-cultural fashion: Focusing on de-Westcentrism cultural discourse and Asian ethnic fashion [*대등적 크로스컬처 패션에 대한 탐색적 연구*]. *Journal of the Korean Society of Costume, 74*(4), 55-73. https://doi.org/10.7233/jksc.2024.74.4.055

Park, S., & Cho, M. (2025). A study on the reproduction of Korean traditional purple dyeing methods using records of redroot gromwell in the late Joseon Dynasty [*조선 후기 문헌에 기록된 자초 염색을 활용한 전통 자색 재현 방법 연구*]. *Journal of the Korean Society of Costume, 75*(6), 42-56. https://doi.org/10.7233/jksc.2025.75.6.042

Sungshin Women's University Museum. (n.d.). *Jobok husu/daedae* [*조복 후수/대대*]. e-Museum. https://www.emuseum.go.kr/detail?relicId=PS0100501500400076500000

Takahashi, S., Yamaguchi, K., & Watanabe, A. (2021). A novel approach to analyze fashion digital library from humanities. In *Towards Open and Trustworthy Digital Societies: 23rd International Conference on Asia-Pacific Digital Libraries (ICADL 2021), Lecture Notes in Computer Science* (Vol. 13133, pp. 179-194). Springer. https://doi.org/10.1007/978-3-030-91669-5_15

Wei, J., Yang, C., Song, X., Lu, Y., Hu, N., Huang, J., ... Le, Q. V. (2024). Long-form factuality in large language models. *Advances in Neural Information Processing Systems, 37,* 80756-80827. https://doi.org/10.52202/079017-2567

Wei, Y., Xu, Y., Wei, X., Yang, S., Zhu, Y., Li, Y., ... Wu, B. (2024). AC-EVAL: Evaluating ancient Chinese language understanding in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 1600-1617). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.87

Wu, W., Wang, H., Li, B., Huang, P., Zhao, X., & Liang, L. (2025). MultiRAG: A knowledge-guided framework for mitigating hallucination in multi-source retrieval augmented generation. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)* (pp. 3070-3083). IEEE. https://doi.org/10.1109/ICDE65448.2025.00230

Xu, L., Lu, L., Liu, M., Song, C., & Wu, L. (2024). Nanjing Yunjin intelligent question-answering system based on knowledge graphs and retrieval augmented generation technology. *Heritage Science, 12,* Article 118. https://doi.org/10.1186/s40494-024-01231-3

Ye, W., Zheng, T., Qi, Y., Zhao, W., Wang, X., Zhao, X., ... Wang, D. (2025). *ICH-Qwen: A large language model towards Chinese intangible cultural heritage.* arXiv. https://doi.org/10.48550/arXiv.2505.23831

Yu, J., Zhang, L., Chen, Z., Pan, F., Wen, M., Yan, Y., ... Lan, Z. (2023). *Quality and quantity: Unveiling a million high-quality images for text-to-image synthesis in fashion design.* arXiv. https://doi.org/10.48550/arXiv.2311.12067

Zucchet, N., Bornschein, J., Chan, S., Lampinen, A., Pascanu, R., & De, S. (2025). How do language models learn facts? Dynamics, curricula and hallucinations. In *Proceedings of the Second Conference on Language Modeling (COLM 2025).* https://openreview.net/forum?id=vBcGnragkr