Interview https://doi.org/10.23287/KJDH.2026.3.1.5

디지털 인문학에 관한 테드 언더우드 교수와의 인터뷰*

An Interview with Professor Ted Underwood on Digital Humanities

테드 언더우드(Underwood, Ted)**
0000-0001-8960-1846

목차

* 이 인터뷰는 번역본과 원문을 순차적으로 수록하였다. 김희진(경북대학교 영어영문학과 부교수, hkim1596@knu.ac.kr)이 2025년 12월 18일 김희진 교수의 연구실에서 인터뷰를 진행하였으며, 인터뷰 질문 준비 및 번역 과정에는 강수진(경북대학교 영어영문학과 BK21 교수)과 노유정(경북대학교 영어영문학과 박사과정)이 참여하였다.
The translated text and original text are presented sequentially. The interview was conducted by Heejin Kim (Associate Professor, Department of English Language and Literature, Kyungpook National University, hkim1596@knu.ac.kr). Sujin Kang (BK21 Professor, Department of English Language and Literature, Kyungpook National University) and Yujeong Noh (Ph. D. Candidate, Department of English Language and Literature, Kyungpook National University) contributed to the preparation of the interview questions and the translation process.

** 1st Author: University of Illinois Urbana-Champaign, School of Information Sciences; Department of English, tunder@illinois.edu

1 인터뷰 번역본

인터뷰에 응해 주셔서 진심으로 감사드립니다. 교수님께서는 일리노이 대학교의 정보과학대학과 영어영문학과에 소속되어 계시며, 전산 문학 연구와 디지털 인문학 분야에서 중요한 기여를 해오셨습니다. 특히 문학사 연구에 계산적 방법을 적용하고, 모델링과 해석의 관계를 재고하는 작업을 통해 디지털 인문학의 방법론적 지평을 확장해 오신 점에 깊이 감사드립니다. 교수님의 답변은 영어와 한국어로 함께 제공되어, 한국의 연구자들이 교수님의 연구와 디지털 인문학의 향방에 대한 통찰과 문제의식을 보다 폭넓게 접할 수 있는 기회를 마련하고자 합니다.

그럼, 이제 본격적으로 질문을 시작하겠습니다.

1.1 멀리서 읽기에서 모델링으로

(1) *교수님의 초기 연구 중 상당 부분은 단어 빈도, 토픽 모델링, 장르 분류기처럼 해석 가능한 특징에 의존했습니다. 반면 최근 프로젝트들은 임베딩, 퍼플렉서티, 딥러닝 표현을 활용하고 계신데요. 이러한 변화가 문학사 연구에 있어 어떠한 개념적 전환을 의미하는지 궁금합니다. 또한 저서 『Distant Horizons』에서 교수님은 결정적인 변화가 단순히 컴퓨터의 도입에 있는 것이 아니라, 모델링의 선택과 그 모델링이 가능케 하는 해석적 태도에 있다고 주장하셨습니다. 지금 돌이켜보았을 때, 독자들이 모델링의 역할에 대해 가장 흔히 오해하고 있는 부분은 무엇이라고 생각하십니까?*

우리가 모델을 구축하는 방식이 통계적 머신러닝에서 표현 학습으로 변화하고 있다는 기술적 질문과, 그 흐름 속에서 '멀리서 읽기'가 어디에 위치하느냐는 다소 다른 질문이 존재합니다. 우선 멀리서 읽기와 기술적 변화의 상관관계에 대해 말씀드리고 싶습니다.

근본적으로 멀리서 읽기의 시대는 저물고 있다고 볼 수 있습니다. 멀리서 읽기라는 아이디어 자체가 인간의 독해가 할 수 있는 일과 컴퓨터가 할 수 있는 일 사이의 노동 분업에 의존해 왔기 때문입니다. 과거에는 컴퓨터가 인간 독자만큼 해내지 못하는 일이 많았고, 반대로 인간 독자는 대규모 작업을 수행할 수 없었습니다. 이러한 분업이 있었기에 꼼꼼하게 읽기(close reading)와 멀리서 읽기 사이의 대조를 논하는 것이 합리적이었습니다.

하지만 거대언어모델(Large Language Models, 이하 LLM)의 등장으로 이러한 대립 구도는 무너지고 있습니다. 이제 전산 모델은 우리가 과거에 꼼꼼하게 읽기라고 규정했던 작업들을 수행할 수 있게 되었습니다. 비록 아직은 인간 독자만큼 뛰어나지 않을지 모르나, 그것이 본질적인 종류의 차이는 아닙니다. 그래서 저는 더 이상 멀리서 읽기라는 용어를 즐겨 쓰지 않습니다. 현재 이 분야를 설명하기에는 '전산 문학 연구(computational literary study)'나 '전산 인문학(computational humanities)'이 더 적절한 표현일 것입니

다. 전산적 접근법과 여타 접근법 사이에는 여전히 차이가 존재하지만, 저는 그것이 더 이상 원격 대 근접이라는 축으로 조직된다고 보지 않습니다.

이와는 별개로, 새로운 모델들이 어떤 가능성을 열어주는가에 대한 질문이 남습니다. 이 모델들이 단순히 인간의 꼼꼼하게 읽기를 복제한다고 말하는 것은 옳지 않습니다. 모델이 인간보다 열등한 영역도 여전히 존재하지만, 과거에는, 심지어 꼼꼼하게 읽기로도 불가능했던 새로운 방식의 대조적 접근법이나 모델 해석법을 가능하게 하기도 합니다. 근접과 원격의 대립은 인간과 전산적 역량 각자의 고유한 강점과 역량에 대한 주목으로 대체될 것이며, 이는 반드시 규모의 문제로만 국한되지는 않을 것입니다.

모델링에 관해 말씀드리자면, 『Distant Horizons』에서 강조한 지점, 즉 우리가 단순히 '측정'하는 것이 아니라 '모델을 구축'하고 있다는 사실을 짚어 주셔서 기쁩니다. 저는 대다수의 인문학자들이 아직 이 점을 충분히 수용하지 못했다고 생각합니다. 현재 우려되는 점은 LLM의 대화형 인터페이스가 제공하는 가시적인 투명성으로 인해, 사람들이 자신이 무엇을 하고 있는지, 즉 '모델이란 무엇인가'에 대해 성찰하지 않기 쉬워졌다는 것입니다. 이제 '모델'이라는 용어는 감각이 마비된 단어가 되어가고 있으며, 사람들은 이를 단순한 컴퓨터 프로그램 정도로 취급합니다. 우리가 모델링에 대해 진지하게 사유해 본 적이 없다는 점, 그리고 여전히 사유가 필요하다는 점을 고려할 때 이는 상당한 리스크라고 할 수 있습니다.

(2) *저 역시 LLM의 등장 이후 멀리서 읽기와 꼼꼼하게 읽기의 경계가 더 이상 견고하지 않다는 점에 동의합니다. 저는 수업 시간에 학생들에게 LLM을 활용한 꼼꼼하게 읽기를 시도해보라고 권합니다. 예를 들어, 셰익스피어의 『헨리 6세 제3부』에는 갇혀 있는 사자의 심상이 등장하는데, 학생들에게 이 심상의 계보를 추적해보라고 과제를 내주었습니다. 학생들은 API를 활용해 셰익스피어의 전체 작품군 내에서 유사한 사자 심상을 탐색했습니다. 만약 우리가 이를 현대 문학이나 이북으로 출간된 방대한 도서와 같이 훨씬 더 거대한 코퍼스를 대상으로 수행할 수 있다면, 이러한 방식의 심상 탐색 접근법이 유용할 것이라고 보십니까?*

네, 전적으로 동의합니다. 말씀하신 내용은 확장된 차원, 즉 확장된 맥락 속에서 이루어지는 꼼꼼하게 읽기처럼 들립니다. 여기서 규모의 확장은 반드시 분석 대상이 되는 구절의 양이 늘어나는 것을 의미하기보다, 그 구절을 해석하기 위해 동원할 수 있는 자원이 방대해진다는 것을 의미합니다. 매우 흥미롭군요. 솔직히 말씀드리면 저는 그 부분까지는 생각지 못했는데, 대단히 흥미로운 접근 방식입니다.

1.2 선형적 시대를 넘어서

(3) *딥러닝 모델은 사전 정의된 범주를 검증하기보다 잠재된 구조를 발견해내는 경우*

가 많습니다. 이러한 특성이 장르나 시대 구분, 혹은 영향 관계에 대한 교수님의 생각에 어떤 변화를 주었나요? 만약 모델이 기존의 해석적 공동체나 제도적 분류 체계와 상충하는 결과를 내놓는다면, 그것을 새로운 발견으로 보십니까? 아니면 교정해야 할 오류가 아니라 그 자체로 보존할 가치가 있는 관점의 불일치로 보십니까?

저는 모델을 하나의 증거로 간주합니다만, 때로는 그것이 특정한 관점에 관한 증거가 되기도 합니다. 멀리서 읽기가 가장 효과적이었던 지점, 그리고 문학에 대한 저의 생각을 가장 강력하게 변화시킨 지점은 시대 구분이나 영향 관계와 같은 시간에 따른 변화의 문제입니다. 우리는 사건이나 불연속적인 시기에 의존하는 변화 모델들이 대체로 옳기보다 틀렸다는 사실을 알게 되었습니다. 변화는 연속적이며, 코호트 대체(cohort replacement)를 통해 이루어집니다. 이는 기존의 문학사가 주로 인용해온 변화 모델과는 사뭇 다른 모습입니다. 이러한 통찰은 부분적으로는 모델링에, 부분적으로는 측정에 의존한 결과이지만, 제가 멀리서 읽기를 통해 얻은 가장 큰 수확 중 하나입니다.

하지만 때로 장르의 안정성이나 불안정성 같은 것을 이해하려 할 때, 우리는 모델에서 인간 독자들의 증언과 일치하지 않는 지점들을 발견하곤 합니다. 인간 독자들이 "이 책들은 SF다"라고 말하고, 우리가 모델을 훈련시켰을 때 모델이 "네, 이 모든 것을 통합하는 무언가를 감지할 수 있습니다"라고 답해준다면 더할 나위 없이 좋겠지요. 하지만 저의 저서 『Distant Horizons』에서 고딕 장르를 대상으로 이를 시도했을 때, 모델은 그리 높은 정확도를 보여주지 못했습니다.

모델이 인간의 라벨링과 일치하지 않을 때 까다로운 점은, 부정적인 결론을 입증하기가 어렵다는 사실입니다. 즉, 특정 작품들이 서로 공통점이 없다는 것을 증명하기란 쉽지 않습니다. 우리가 증명한 것이라고는 단지 모델에 제공한 증거의 종류나 모델이 도출할 수 있는 변별력의 유형이 공통점을 찾아내기에 충분하지 않았다는 사실뿐일 수도 있기 때문입니다. 따라서 신중해야 합니다. 이러한 방법론으로는 긍정적인 결론을 내리기는 쉽지만, 부정적인 결론을 도출하기는 대단히 어렵습니다. 이것이 바로 멀리서 읽기가 변화를 포착하는 데는 탁월했으나, 장르에 대한 통찰을 만들어내는 데는 항상 효과적이지는 않았던 이유 중 하나입니다. 연속성은 감지할 수 있고 연속성을 찾을 수 없는 사례를 기술할 수는 있지만, 연속성의 '부재' 자체를 증명하는 것은 어려운 일입니다.

(4) *제가 듣기로 누군가는 '침묵'의 문제를 제기하더군요. 즉, 증거의 부재가 세계에 대한 왜곡된 이미지를 재생산할 수 있다는 지적입니다. 전산적 방법론은 이러한 아카이브의 침묵(archival silence) 문제를 어떻게 다루고 있습니까?*

그것은 또 다른 차원의 문제들입니다. 아카이브의 침묵을 전산적으로 어떻게 처리해야 할지 저로서는 아직 확답을 드리기 어렵습니다. 로렌 클라인(Lauren Klein)이 그 분야를 연구해왔고 관련 논문들도 쓴 것으로 알고 있습니다만, 그러한 연구들조차 여전히 인간 해석자에게 크게 의존하고 있습니다. 전산적 증거를 확보할 수는 있겠지만, 침묵이나 증

거의 부재가 발생하는 지점에서는 현재로선 인간이 무언가를 기여해야만 합니다. 이를 보조할 수 있는 전산적 방법들을 구상해볼 수는 있겠으나, 현재 제가 구체적으로 알고 있는 바는 없습니다.

1.3 데이터 소스와 해석 가능성

(5) *아까 차 안에서 출처를 추적하기 위해 지식 그래프(knowledge graph)를 활용하는 방안에 대해 이야기를 나누었지요. 일단 어떤 데이터가 모델 내부에 들어가면 수치화되어 버리고, 그 기원은 소실되기 때문입니다. 모델을 통해 발견한 결과물들의 증거적 지위를 높일 수 있는 방법이 있을까요?*

네, 지식 그래프는 매우 흥미로운 접근 방식입니다. 언어 모델은 많은 것을 불투명하게 만드는데, 어떤 소스들이 사용되었는지가 그 불투명한 요소 중 하나입니다. 지식 그래프가 그 지점에서 도움을 줄 수 있을 것입니다.

또 다른 방법으로는, 비록 노동 집약적이긴 하지만 훈련 세트에서 특정 항목들을 제거해보는 방식이 있습니다. 완전히 다른 훈련 데이터로 모델을 처음부터 다시 훈련시키는 것은 극도로 힘든 일이겠지만, '체크포인트'를 활용하는 방법은 가능할 것입니다. 즉, 어느 지점까지는 동일하게 훈련시킨 뒤 가지를 쳐서, 서로 다른 데이터를 가진 여러 모델로 분화시켜 훈련하는 것이죠. 이를 통해 어떤 소스가 결과에 차이를 만들어내는지 추론할 수 있을 것입니다. 이것이 바로 우리가 역사적 언어 모델을 통해 시도하고자 하는 일 중 하나입니다.

더 넓게 보면, 저는 기계론적 해석 가능성을 통해 LLM과 신경망 모델의 내부를 '열어젖혀야' 한다고 생각합니다. 서로 다른 프롬프트를 주었을 때, 혹은 서로 다른 데이터로 훈련된 유사한 두 모델을 비교할 때, 모델의 추론 과정이나 활성화, 가중치가 어떻게 변하는지 질문을 던지는 것이죠. 과거의 통계적 머신 러닝에서는 내부 프로세스를 조사하기가 쉽지 않았기에 이러한 지렛대를 활용하기 어려웠습니다. 흔히 언어 모델이 더 불투명하다고들 하지만, 어떤 면에서는 덜 불투명하기도 합니다. 이제 막 열리기 시작한 다양한 조사 방식들이 있으며, 이는 충분히 시도해 볼 가치가 있습니다.

또한, 추론 모델에서 나타나는 추론 흔적을 살펴보는 것도 유의미합니다. 이것이 언제나 특정한 소스를 지목해 주지는 않겠지만, 중요한 단서를 제공해 줄 수는 있을 것입니다.

(6) *어제 김바로 교수님께서도 언급하셨듯이, 해석 가능성을 높이기 위해 서로 다른 모델과 임베딩, 랭킹 시스템을 결합하는 '검색 증강(retrieval-augmented)' 접근법이 현재 매우 각광받고 있습니다.*

그렇습니다. 하지만 모델이 무엇을 바탕으로 훈련되었는지 알고 이를 제어할 수 있기 전까지는, 그 어떤 방법론도 온전한 의미를 갖기 어렵습니다.

1.4 모델 평가와 해석 전략

(7) *텍스트 간 거리와 혁신에 관한 연구에서 교수님은 토픽 모델링, 임베딩, 그리고 퍼플렉서티 방식의 측정 도구들을 비교하셨습니다. 역사학적 도구로서 딥러닝이 가진 강점과 한계에 대해 무엇을 배우셨나요? 또한 교수님은 신경망 모델이 어휘 기반 모델보다 체계적으로 더 우수한 성능을 보였다는 증거가 없다고 보고하셨습니다. 디지털 인문학 연구자들은 베이스라인, 해석 가능성, 혹은 성능의 정의와 관련하여 이 결과로부터 무엇을 추론해야 할까요?*

간단히 답하자면, 단지 최신 기술이라는 이유만으로 최첨단 방법론이 반드시 필요하다고 자동적으로 가정해서는 안 된다는 것입니다. 어떤 질문들에 대해서는 오래된 방식, 심지어 아주 단순한 방식들이 여전히 유효합니다. 만약 정확도가 크게 떨어지지 않으면서도 방법론이 더 단순하고 특정 측면에서 해석 가능성이 높다면, 우리는 당연히 그 방식을 택해야 합니다.

현 상황에서는 기술의 변화 속도가 워낙 빠르다 보니, 임베딩을 사용할 수 있는 상황에서 어휘적 증거나 토픽 모델링으로 논증을 펼치는 것을 주저하는 경향이 있어 어려움이 따르기도 합니다. 하지만 많은 경우 우리는 필요 이상으로 연구 과정을 복잡하게 만듭니다. 물론 항상 그렇다는 것은 아닙니다. 저 역시 앞으로 LLM을 활용한 연구를 수행할 것입니다. 다만 제가 바라는 전산 인문학의 방법론적 미래는 단순한 모델이 지닌 가치를 명확히 인식하는 모습입니다.

단순한 모델이 중요한 또 다른 이유는 저작권 데이터 때문입니다. 저작권 문제로 인해 전체 텍스트가 아닌 단어 빈도 정보만 가질 수 있는 경우가 많습니다. 현재 우리는 최신 모델을 쓸 수 없다는 이유로 저작권 데이터 연구를 거부하며 스스로를 제약하곤 하는데, 사실 '단어 주머니(bag-of-words)' 방식의 연구로 할 수 있는 것들을 기꺼이 수행해야 합니다.

또한 기술적으로는 구식이 되었더라도 해석 가능성 덕분에 여전히 유용한 도구들이 있습니다. 예를 들어 '도큐스코프(DocuScope)' 같은 도구는 기술적으로는 뒤처졌을지 몰라도, 고도의 해석 가능성을 제공한다는 점 때문에 여전히 유용합니다.

(8) *매우 인상적인 발견 중 하나는 문서를 구성하는 여러 구절 중 가장 미래지향적인 대목들을 통해 문서를 재현했을 때 그 신호가 더 강력해진다는 점입니다. 이러한 사실이 문학 모델링, 그리고 우리가 역사적으로 유의미하다고 간주하는 '단위'의 설정에 어떤 시사점을 주는지요? 만약 혁신이 특정 부분에 집중되어 있다면, 검증 방법으로서의 가까이*

서 읽기에는 어떤 변화가 뒤따라야 할까요? 우리는 어디를 읽어야 할지에 관한 새로운 프로토콜이 필요한 것일까요?

대단히 흥미로운 지적입니다. 특히 마지막 질문은 저희가 논문에서 실제로 언급했던 수준 너머의 가능성을 보여주네요. 논문에서 저희가 주장한 바는, 저자들이 작품의 아주 일부분에서만 혁신을 시도하며, 따라서 작품의 모든 부분이 혁신적일 것이라 기대하는 것은 오류라는 점입니다.

꼼꼼하게 읽기의 관점에서 보자면, 현재 우리는 작품의 어느 부분이 가장 흥미로운지, 혹은 어디가 면밀히 주목할 가치가 있는지 결정할 수 있는 객관적인 기준을 갖고 있지 않습니다. 하지만 잠재적으로 우리는 텍스트의 어느 부분이 시대의 흐름을 앞서가고 있는지 식별해낼 수 있습니다. 만약 어떤 세부 묘사가 유의미하거나 변혁적이라고 주장하고 싶다면, 아마도 그러한 혁신적인 대목들에서 증거를 찾아야 할 것입니다. 제가 당장 어떤 규칙을 만들고 싶은 것은 아니지만, 적어도 우리가 어디를 유심히 살펴봐야 할지 알려주는 유용한 발견적 방법이 될 수는 있을 것입니다.

(9) *추리 소설 사례 연구에서 교수님은 고차원 모델이 우리에게 명쾌한 정의를 제공할 의무는 없다고 경고하셨습니다. 모델을 '열어젖힐' 수는 있지만, 여전히 과잉 결정된 상태일 것이라는 점이지요. 그렇다면 문학사 연구에서 이러한 모델에 대한 수용 가능한 설명이란 무엇을 의미할까요? 교수님은 극히 흔한 단어들이나 물음표 같은 문장 부호만으로도 추리 소설을 놀라울 정도로 잘 분류할 수 있음을 보여주셨습니다. 이러한 결과가 장르를 어떤 '주제'라기보다, '담론적 태도'로 바라보게 하는 계기가 되었나요?*

제 대답은 신중할 수밖에 없습니다. 제가 거기서 살펴본 모델들은 주로 예측적 기능을 수행할 뿐, 반드시 '설명적 모델'인 것은 아니기 때문입니다. 모델의 특징들을 어떻게 설명으로 전환할지에 대한 규칙이 있는지도 의문이며, 특정 요소의 중요성이 우리에게 장르에 대한 새로운 이론을 제공한다고 확신하기도 어렵습니다.

장르에 대한 저의 이론은 다분히 느슨하고 사회적입니다. 그것은 독자와 작가가 작품을 어떻게 범주화하느냐의 문제와 관련이 있습니다. 예측 모델을 바탕으로 장르에 대한 설명적 이론을 전개하는 것에는 신중한 입장입니다. 그러고 싶은 유혹이 들기도 하지만, 그러한 모델들은 변수가 너무 많기 때문에 인과적으로 읽어낼 수가 없습니다.

물론 우리가 영원히 이러한 신중한 자세에만 머물러야 한다고 생각하지는 않습니다. 저는 언어 모델을 해석하는 가능성—서로 다른 층위를 살펴보고, 추론 흔적을 추적하며, 모델들을 대조하는 방식—에 깊은 흥미를 느끼고 있습니다. 그러한 증거로부터 어떻게 '설명'에 도달할 수 있을지는 아직 알 수 없지만, 흥미로운 가능성의 문이 열려 있는 것만은 분명합니다.

1.5 표집과 역사적 논증

(10) *한 세트의 판단 기준으로 훈련하고 다른 세트로 테스트하는 교수님의 방식은 장르를 여러 라벨링 관행에 걸쳐 분산된 것으로 취급합니다. 이러한 전이의 실패는 무엇을 드러낼까요? 경계에 선 사례들인가요, 아니면 시대적 변화, 혹은 메타데이터에 내재된 제도적 역사인가요? 더 넓게는, 디지털 인문학 연구자들이 목록 시스템을 교정해야 할 '노이즈가 섞인 라벨'로 보아야 할까요, 아니면 그 자체로 역사적 주체성을 지닌 '해석적 공동체'로 보아야 할까요?*

지금까지 저는 그것들을 해석적 공동체로 취급하는 쪽에 무게를 두어 왔습니다. 장르가 텍스트 내에 실재하는 어떤 것이어서, 우리가 장르에 대한 인간의 행동을 교정할 권리를 갖게 된다고는 생각하지 않습니다. 저는 장르가 인간의 행동 양식 안에 구속되어 있다고 봅니다.

하지만 라벨의 종류도 다양합니다. 목록 시스템은 사회적 증거의 한 종류일 뿐입니다. 학술적 서지 자료도 있고, 굿리즈(Goodreads)나 중국의 더우반(Douban) 같은 팬 커뮤니티 역시 또 다른 종류의 증거가 됩니다.

예를 들어, 독자들이 생성한 리뷰나 라벨이 도서관의 목록 시스템이나 학계의 서지 자료보다 더 정확하고 일관된 장르 모델을 도출한다는 연구 결과를 상상해볼 수 있습니다. 그리고 독자 생성 라벨은 수많은 사람으로부터 나오기에, 독자들이 자신들이 무엇에 대해 말하고 있는지 정확히 알고 있다고 추정할 근거가 될 수 있지요. 하지만 여전히 신중해야 합니다. 반드시 텍스트 모델을 인간의 증언보다 우선시해야 할 이유는 없기 때문입니다. 텍스트 모델이 수많은 인간의 증언과 일치할 때, 비로소 그 결과는 시사점을 갖게 됩니다.

(11) *데이터셋 연구에서 교수님은 표집 틀이 달라도 광범위한 경향성은 안정적으로 유지될 수 있으며, '단 하나의 올바른 표본'에 집착하는 논쟁은 이미 증거가 강력하게 뒷받침하고 있는 사실을 놓치게 만들 수 있다고 주장하셨습니다. 역사적 주장을 정당화하기 위해 어느 정도의 안정성이 확보되어야 '충분하다'고 판단하십니까? 또한 위상, 재판 발행, 정전 형성 등과 같은 가치 평가가 주된 분석 대상이 아닐 때, 이를 배제해야 하는 가장 강력한 이유는 무엇입니까? 그리고 그러한 배제가 오히려 실책이 되는 경우는 언제입니까?*

안정성에 관해서는 정량적인 답변이 가능합니다. "이러한 방식으로 재표집을 수행했을 때, 타겟 값에 이 정도의 변화가 발생한다. 우리가 보고하는 효과 크기, 가령 시간에 따른 유의미한 변화는 이 정도이다"라고 말할 수 있지요. 이 두 가지를 비교함으로써, 통계적 분석을 통해 해당 효과가 단순한 변동성에 의해 설명될 가능성이 낮음을 입증할 수

있습니다. 이는 신호의 어느 정도가 우연에 의한 것인지를 따지는 표준적인 접근법입니다.

위상의 문제는 더 까다롭습니다. 당신이 펼치려는 주장의 성격에 따라 어떤 종류의 표본이 필요한지가 달라지기 때문입니다. 독자에 관한 주장을 하려는 것인지, 혹은 변화에 관한 주장을 하려는 것인지에 따라 다릅니다. 널리 읽힌 작품이 변화를 일으킬 가능성이 더 크다는 이론을 세울 수도 있겠지만, 그러한 이론들 역시 검증이 가능합니다. 변화에 관한 논증을 반드시 위상이 높은 작품들에 근거해야 할 선험적인 이유가 있는지는 모르겠습니다. 아마도 위상 높은 작품들이 실제로 더 영향력이 있다는 증거가 얼마나 강력한지를 먼저 테스트해 보아야 할 것입니다.

실제로 우리는 시대를 앞서간 작품들이 특정 종류의 위상 신호와 일치한다는 증거를 가지고 있으며, 위상의 종류에 따라 그 영향력이 다르다는 사실도 알고 있습니다. 예를 들어, 평단에서 널리 리뷰되는 것과 전위적이라고 인식되는 것 중 무엇이 텍스트적으로 시대를 앞서가는 것과 더 밀접하게 상관관계가 있는지 질문을 던져볼 수 있을 것입니다.

1.6 LLM 시대의 문해력과 실천적 전략

(12) *챗GPT와 같은 시스템의 등장이 전산적 방법론, 문학사, 혹은 이론을 가르치는 교수님의 방식에 어떤 변화를 가져왔습니까? 인문학적 맥락에서 학생들이 LLM을 책임감 있게 활용하기 위해, 예를 들어 훈련 데이터, 평가, 불확실성, 그리고 제도적 유인 구조 등에 관하여 이제 어떤 종류의 개념적 문해력을 갖추어야 한다고 보십니까?*

우리 중 누구도 이 분야의 전문가라고 할 수는 없습니다. 이를 고민할 시간이 고작 3년 남짓이었으니까요. 제 생각에 학생들은 이제 훨씬 더 빠르게 작업을 수행할 수 있게 되었습니다. 예전 같으면 오랜 시간이 걸렸을 일들을 더 빨리 배우고 스스로 익힐 수 있게 되었지요.

하지만 함정도 존재합니다. 무언가를 빠르게 독학하다 보면, 그 방법론의 한계나 스스로를 속이게 되는 방식에 대해서는 미처 알지 못할 수 있습니다. 이는 학생뿐만 아니라 교수진에게도 해당되는 이야기입니다. 따라서 주의가 필요합니다.

또한 다양한 관점을 확보하는 것의 가치에 대해서도 성찰이 필요합니다. 언어 모델로부터는 “이 방법이 효과적일 것”이라는 식의 단일한 관점을 얻게 되지만, 거기에는 맹점이 있을 수 있습니다. 따라서 인간과도 상담해야 합니다. 현재의 언어 모델들은 서로 유사한 맹점을 공유하기 쉽기 때문입니다. 해당 방법론의 함정을 잘 알고 있는 사람의 조언을 들을 필요가 있는 것이지요.

김 교수님의 경험은 어떠신가요? 학생들이 이를 활용해 스스로 무언가를 배우고 있다고 느끼시나요?

(13) *네, 저는 학생들에게 모델을 사용하게 하고 그 대화 로그를 제출하도록 합니다. 결과물이 단순히 LLM을 사용한 수준보다 더 낫기를 기대하면서요. 저는 셰익스피어 판본 편집을 가르치는데, 최근 한 작품의 텍스트적 상태에 관한 학계의 합의에 이의를 제기하는 논문을 쓴 적이 있습니다. 제 입장은 아직 모델의 데이터에 반영되지 않았기 때문에, 모델은 과거의 합의된 내용만을 학습한 대로 제 의견에 반대하는 논거를 펼치는 경향이 있더군요.*

그것이 바로 리스크를 보여주는 아주 좋은 사례입니다. 합의된 견해가 균질화되고, 그로 인해 변화를 만들어내기가 어려워지는 것이죠.

(14) *학생들은 더 많은 증거, 즉 더 최신의 논문들을 컨텍스트 윈도우에 추가하여 모델이 다르게 생각하도록 유도합니다. 그 결과는 꽤 흥미롭습니다.*

학생들이 그렇게 해서 모델이 다르게 생각하도록 만드는 것이 가능한가요?

(15) *네, 그렇습니다.*

그렇다면 장기적으로 우리는 현재 언어 모델에서 얻을 수 있는 것보다 더 다양한 견해의 차이를 이끌어낼 수 있는 방법이 필요합니다. 이는 부분적으로는 모델 자체를 바꾸는 문제이기도 하고, 동시에 인간이 맹점과 성급한 합의를 경계하도록 가르치는 문제이기도 합니다.

학생들이 더 빨리 배우고는 있지만, 우리는 맹점에 대해 우려해야 합니다. 그들이 얻는 첫 번째 답변은 대개 B 수준의 답변입니다. A를 얻기 위해서는 더 많은 입력값을 넣거나, 편집을 하거나, 관점을 다양화하는 등의 추가적인 노력이 필요합니다.

궁극적으로 우리는 사람들에게 대안적인 관점을 찾는 법을 가르칠 프로토콜, 즉 기술이 필요할지도 모릅니다. “이 사안에 대한 의견의 스펙트럼은 어떠한가?”라는 질문을 던질 줄 알아야 한다는 것이죠. 저는 단 하나의 정답이 아니라, 우리가 모르는 부분을 포함하여 잠재적인 관점들이 펼쳐진 지도를 원합니다.

1.7 시대착오, 저자성, 그리고 이론

(16) *언어 모델이 시대착오 없이 과거를 재현할 수 있는지에 관한 연구에서, 교수님은 이 문제가 단순한 기술적 결함이 아니라고 주장하셨습니다. 무엇이 시대착오를 LLM 기반 인문학 연구의 이토록 난해한 과제로 만드는 것입니까? 미세 조정(fine-tuning)이나 역사적 코퍼스로의 제한이 이 문제를 유의미하게 해결할 수 있을까요, 아니면 단지 인식론적 리스크를 다른 곳으로 옮겨놓는 것에 불과할까요?*

여기에는 기술적인 차원이 분명히 존재합니다. 이는 '관점의 균질성' 혹은 이른바 '모드 붕괴(mode collapse)'라 불리는 훈련 데이터에 의해 지배되는 단일한 시각과 관련이 있습니다. 시대착오와 관련하여 발생하는 기술적 문제는 모델이 '망각'할 수 없다는 점입니다. 만약 모델이 현재를 목격했다면, 그것이 존재한다는 사실을 잊을 수 없습니다. 물론 이에 대해서는 기술적인 해결책이 있을 수 있습니다. 이를 모드 붕괴로 파악하는 것 역시 기술적 수정을 통해 모델이 가장 일반적인 답변 외부를 탐색하도록 보장할 방법이 있음을 시사합니다.

하지만 기술 그 이상의 문제는 바로 해석학적 차원에 있습니다. 여기에는 언제나 순환성이 존재합니다. 우리의 관점은 우리가 이해하고자 하는 대상에 의해 형성되며, 이는 객관성에 관한 일부 이론들이 요구하는 것처럼 우리가 대상으로부터 완전히 분리될 수 없음을 의미합니다. 이 문제는 사라지지 않을 것입니다. 그것은 본질적인 문제이기에 기술적인 해결책 또한 존재하지 않습니다. 대신 우리는 이 지점에 대해 끊임없이 성찰해야 합니다.

(17) *저작 「이론의 경험적 승리(The Empirical Triumph of Theory)」에서 교수님은 LLM이 20세기 의미 이론들을 극적으로 재현하고 있다고 제안하셨습니다. 또한 지시어 튜닝이 일종의 가변적인 '저자 기능'을 재구축하는 방식에 대해서도 논하셨지요. 이러한 통찰은 현재 전산 인문학에서 행위성, 의도, 그리고 해석을 논하는 방식에 어떤 함의를 가집니까? 만약 전산 인문학의 다음 단계가 해석과 거버넌스에 의해 정의된다면, 학술지나 연구소, 대학원 교육 과정에서 가장 먼저 채택해야 할 증거, 공시, 그리고 평가의 규범은 무엇이라고 보십니까?*

대단히 방대한 질문들입니다. LLM이 던지는 과제는 전산 인문학이라는 영역보다 훨씬 거대합니다. 인간의 행위성과 지능에 관한 우리의 관념은 그간 주체 중심적 모델에 기반해 왔습니다. 이는 과거 구조주의와 포스트 구조주의에 의해 도전받은 바 있으며, 이제 LLM이 이를 다시 한번 위협하고 있습니다. 이는 우리 전산 인문학 연구자들이 고민해야 할 지점인 동시에, 훨씬 더 큰 사회적 도전의 일부이기도 합니다. 즉, '언어가 인간을 통해 말한다'는 통찰을 사람들이 받아들이기 시작할 때 어떤 일이 벌어질 것인가에 관한 문제입니다. 많은 이들이 이를 공포스럽게 느끼기도 하지요.

증거, 공시, 평가 규범에 관해 말씀드리자면, 저는 코드와 데이터를 공유하는 오픈 사이언스를 지지합니다만, AI의 등장으로 이 문제가 더욱 복잡해졌습니다. 현재 학술 기관들은 AI를 활용한 연구에서 어떤 방식의 감사의 글이나 출처 명기가 적절한지에 대해 활발히 논의 중입니다. 저 역시 단 하나의 정답을 가지고 있지는 않습니다. 이는 우리 공동체가 함께 고민하며 만들어가야 할 숙제입니다.

(18) *마지막으로 덧붙이고 싶은 말씀이 있으신가요?*

아닙니다. 질문들이 정말 철저하고 훌륭했습니다. 좋은 인터뷰를 할 때마다 느끼는 점인데, 제가 그동안 해온 일들을 스스로 얼마나 많이 잊고 있었는지 새삼 깨닫게 됩니다. 제 작업을 통독한 누군가가 저보다 제 연구를 더 잘 알고 있을 때가 있는데, 우리의 기억은 모든 것을 담아두지 못하지만 논문은 그것을 기록하고 있기 때문입니다.

이것이 바로 인터뷰 과정이 흥미로운 이유입니다. 개인의 주관성, 즉 개별 의식의 한계를 발견하게 되고, '쓰기'라는 행위가 인간의 뇌를 얼마나 초과하는지 알게 됩니다. 시간이 흐르면서 우리는 수많은 일을 하지만 그 전부를 기억할 수는 없습니다. 지금의 나는 그때와는 다른 사람이니까요. 그래서 이런 인터뷰가 즐겁습니다. "아, 맞다. 내가 그런 생각을 했었지"라고 다시금 발견하게 되니까요.

(19) *저는 거의 모든 수업에서 교수님의 저작 「이론의 경험적 승리」를 가르칩니다. 학생들이 AI 시대의 주관성과 저자성에 대해 생각하는 데 큰 도움이 되기 때문입니다. LLM이 등장하기 이전에도 바르트(Roland Barthes)와 푸코(Michelle Foucault)는 중요했지만, 교수님 논문의 서두는 그 가치를 정말 명쾌하게 밝혀줍니다.*

언어 모델이 학생들이 그러한 근본적인 사고방식을 체득하는 데 도움이 되기를 바랍니다. 1970년대에는 그저 '화려한 프랑스식 단어들'로 치부되어 흘려보냈을 이론들이, 이제는 매우 구체적인 현실로 다가오고 있으니까요.

2 Interview in English

Thank you so much for agreeing to this interview. As a professor affiliated with the School of Information Sciences and the Department of English at the University of Illinois, you have made significant contributions to the fields of computational literary studies and the digital humanities. In particular, I am deeply grateful for how you have expanded the methodological horizons of the digital humanities by applying computational methods to literary history and rethinking the relationship between modeling and interpretation.

Your responses will be provided in both English and Korean, offering Korean researchers an opportunity to engage more broadly with your insights and critical inquiries regarding the current state and future direction of the digital humanities.

With that, let us begin with the questions.

2.1 From Distant Reading to Modeling

(1) *Much of your earlier work relied on interpretable features—word frequencies, topic models, genre classifiers. More recent projects use embeddings, perplexity, and deep learning representations. What conceptual shift does that represent for literary-historical research? And in Distant Horizons, you argue the decisive change is not just "computers," but modeling choices and the interpretive stance they enable. Looking back now, what do you think readers most often misunderstand about what modeling is doing?*

There's a technical question about what's changing in how we build models—from statistical machine learning to representation learning. And there's a somewhat different question about where "distant reading" fits in. Let me start by talking about distant reading and how these technical changes are related to it.

Fundamentally, the era of distant reading may be drawing to a close because the idea relied on a division of labor between things we could do with human reading and things computers could do. Computers couldn't do many things human readers could do, and human readers couldn't work at scale. That division made it reasonable to talk about an opposition between close and distant reading.

But with large language models, that opposition is breaking down. Computational models can now do things we used to describe as close reading. They may not do them as well as human readers yet, but it's not a difference of kind. So I don't use the term "distant reading" as much anymore. "Computational literary study" or "computational humanities" is closer to how I'd describe the field now. There is still a difference between computational approaches and other approaches, but I'm not sure it's organized by distant versus close anymore.

Separate from that, there's the question of what possibilities these new models open up. It wouldn't be right to say they just replicate human close reading. There are still things they're not as good at as human beings. But there are also things they enable that we didn't used to be able to do, and that maybe even close reading didn't allow—different kinds of contrastive approaches, and different approaches to

model interpretation. The opposition of close and distant may be replaced by a focus on different human strengths and different computational capacities, but I'm not sure it's organized around scale.

On modeling: I'm glad you brought out that point from Distant Horizons—that it's important to understand we're building models, not just measuring. I don't think most humanists have absorbed that yet. What's risky now is that the appearance of transparency created by the conversational interface of large language models may make it easy for people not to reflect on what they're doing—on what a model is. The term "model" is becoming numb; people treat it as just a computer program. That's a risk, because we never really thought it through, and we still need to.

(2) *I agree the boundary between distant and close reading isn't solid anymore after LLMs. In my class, I ask students to do close reading with LLMs. For example, there's imagery of a pent-up lion in Shakespeare's Henry VI, Part 3, and I asked them to trace the genealogy of that imagery. Students used the API to search across Shakespeare's canon for similar lion imagery. If we could do that across much larger corpora—say, modern literature, or even most published books available as ebooks—do you think that kind of "image search" approach might be useful?*

Yes, absolutely. What you're describing sounds like close reading with an enlarged dimension—an enlarged context. The increase of scale isn't necessarily in the passages being analyzed, but in what can be brought to bear on them. That's fascinating. To be honest, I hadn't thought about that, but it's an interesting way to approach it.

2.2 Beyond Linear Literary History

(3) *Deep learning models often discover latent structure rather than test predefined categories. How has that changed the way you think about genre, periodization, or influence? When a model disagrees with an interpretive community—or with institutional labels—how do you decide whether you've found a discovery, or a misalignment of perspectives worth preserving rather than "fixing"?*

I think about models as evidence, but sometimes it's evidence about a perspective.

Where distant reading has been most effective, and where it has changed the way I think about literature most powerfully, is in questions about change over time—periodization and influence. We've learned that models of change dependent on events and discrete periods were largely wrong, more wrong than right. Change is continuous, and it works through cohort replacement. It's a different model of change than our literary histories tend to invoke. That has depended partly on modeling and partly on measurement, but it's one of the biggest things I take away from distant reading.

But sometimes when we try to understand something like the stability or instability of a genre, we see things in the model that don't confirm what human readers say. It's great when human readers say "these books are science fiction," we train a model, and the model says, "Yes, I can detect something that unifies all this." But in Distant Horizons, I tried that with the Gothic, and the model didn't have very high accuracy.

What's tricky when a model disagrees with human labels is that it's hard to prove a negative. It's hard

to prove works don't have something in common. All we've proven is that the kind of evidence we gave the model, and the kinds of discriminations it could draw, weren't sufficient to find it. So you have to be cautious. It's easy to draw positive conclusions, but it's really hard to draw negative conclusions from that kind of method. That's part of why distant reading was really good with change, but not always as effective at creating insights about genre. You can detect continuity, and you can describe cases where you can't find continuity, but it's hard to prove its absence.

(4) *I heard someone raise the issue of "silence"—that absence of evidence can reproduce a distorted image of the world. How do computational methods deal with archival silence?*

That's another set of problems. I don't know quite how we deal with archival silence computationally. I know Lauren Klein has worked on that and has articles about it. But they still depend a lot on the human interpreter. You may have computational evidence, but when there's silence or an absence of evidence, right now that's where the human has to contribute something. I can imagine computational methods that would help us with that, but I don't know them right now.

2.3 Data Sources and Interpretability

(5) *We talked in the car about using a knowledge graph to trace sources. Because once something is in the model it becomes numbers and the origin is lost. Is there any way to increase the evidential status of what we find through models?*

Yes, knowledge graphs are an interesting approach. Language models make a lot opaque, and one of the opaque things is what sources are being used. Knowledge graphs might help with that.

Another way, although it's labor-intensive, is to remove things from the training set. If you trained models from scratch with completely different training data, that would be extremely labor-intensive, but there might be ways of checkpointing—training up to a point and then branching and training multiple models with different data. That might give us a way to reason about which sources make a difference. That's one of the things we hope to do with historical language models.

More broadly, I'm increasingly thinking we need to "open up" large language models and neural models using mechanistic interpretability—asking what changes about the model's reasoning process, activations, and weights when we give a different prompt, or compare two similar models trained on different data. That kind of leverage wasn't really available with statistical machine learning, because internal processes weren't as easy to probe. With language models, people say they're more opaque, but in some ways they're less opaque—there are different ways of probing them that are beginning to open up and are worth doing.

Another thing worth looking at is the reasoning trace in a reasoning model. It won't always point to a particular source, but it might give clues.

(6) *As Baro Kim mentioned yesterday, retrieval-augmented approaches are very popular—using different models, embeddings, and ranking systems—to improve interpretability.*

Yes. None of these methods make much sense until we know what the model was trained on and can control it.

2.4 Evaluation and Interpretive Strategy

(7) *In work on textual distance and innovation you compare topic models, embeddings, and perplexity-style measures. What have you learned about the strengths and limits of deep learning as a historical instrument? You also report no evidence that neural models systematically outperformed lexical models. What should DH infer from that—about baselines, interpretability, or what "performance" means?*

The simple answer is that we shouldn't automatically assume we need state-of-the-art methods just because they're new. For some questions, older methods—even very simple methods—work. And if accuracy isn't much lower, and the method is simpler and more interpretable in certain ways, we should prefer it.

That's been hard, because things move quickly and there's hesitation to make an argument with lexical evidence or topic models when embeddings are available. But in many cases we complicate our lives more than we need to. Not always—I'm certainly going to do work with large language models. But I'd like to see a methodological future for computational humanities where we're aware of the value of simple models.

Part of the reason simple models matter is that with copyrighted data we may only have word-count information and not full text. Right now we handicap ourselves by refusing to work on copyrighted data if we can't use the latest model, when we should often do what we can with bag-of-words research.

There are also tools that are technologically outdated but remain useful because of interpretability. DocuScope, for example—outdated in terms of technology, but useful precisely because it can be highly interpretable.

(8) *One striking finding is that the signal strengthens when you represent a document via its most forward-looking passages. Why does that matter for literary modeling and for the kinds of "units" we treat as historically meaningful? If innovation is concentrated, what follows for close reading as a validation method—do we need new protocols for where to read?*

That's really interesting—especially the final part, because it pushes beyond what we actually said in the paper. What we said is that the result suggests authors are making innovations in a small part of the work, and it's a mistake to expect a work to be innovative throughout.

For close reading: right now, when we close read a piece, we don't have a way to decide which part is most interesting or most worthy of close attention. Potentially—hypothetically—we could identify which parts of a text are ahead of the curve. If you want to argue that a detail is significant or transformative, maybe it should come from those parts. I don't want to make rules, but it could at least be a heuristic to tell us where to look.

(9) *In your detective-fiction case study, you caution that high-dimensional models don't owe us crisp definitions; they can be "cracked open," but they'll still be overdetermined. What counts as an acceptable explanation of such a model in literary history? You show that extremely common words—and punctuation like question marks—can classify detective fiction surprisingly well. Does that push you toward genre as discursive stance (questioning, inference) rather than topic?*

My answer is cautious: the models I'm looking at there mostly have a predictive function. They're not necessarily explanatory models. I'm not sure there's a rule for turning them into an explanation, or that we can confidently say the importance of certain features gives us a new theory of genre.

My theory of genre is loose and social. It has to do with how readers and writers categorize works. I'm hesitant to develop an explanatory theory of genre based on predictive models. It's tempting, but those models can't really be read causally because there are too many variables.

I don't think we have to stay in that cautious posture forever. I'm intrigued by possibilities for interpreting language models—looking at different layers, reasoning traces, and contrasting models. I don't know how we would get from that evidence to explanation, but there's an intriguing open door there.

2.5 Sampling and Historical Claims

(10) *Your method of training on one set of judgments and testing on another treats genre as distributed across labeling practices. What do failures of transfer reveal—boundary cases, drift, or institutional history embedded in metadata? More broadly: how should DH scholars treat cataloging systems—as noisy labels to correct, or as interpretive communities with historical agency?*

So far I've leaned toward treating them as interpretive communities. I don't think we know genre is something "present in the text" in a way that gives us the right to correct human behavior about genre. I think of genre as bounded in human behavior.

But there are different kinds of labels. Cataloging systems are one kind of social evidence. Scholarly bibliographies are another. Fan communities—Goodreads, and Douban in China—are different kinds of evidence.

You could imagine research showing, for instance, that reader-generated reviews and labels yield genre models that are more accurate and coherent than what we get from cataloging systems or scholarly bibliographies. And because reader-generated labels come from many readers, that might give a reason to suspect readers know what they're talking about. But you still have to be cautious: there's no reason necessarily to prefer textual models over human testimony. If the textual models line up with a lot of human testimony, then it becomes suggestive.

(11) *In your dataset work, you argue broad trends can be stable across different sampling frames, and debates about "the one right sample" can miss what the evidence can already robustly support. How do you decide what level of stability is "enough" to justify a historical claim? When valuation*

(prestige, reprinting, canonicity) is not the target, what are the strongest reasons to bracket it—and when does bracketing become a mistake.

On stability: there are quantitative ways to answer that. You can say, when we resample in these ways, we get this much change in our target. The effect size we're reporting—say, a significant change across time—is this big. Compare those two, and with statistical work you can say it's unlikely the effect is explained by variation. That's a standard approach: how much of the signal is likely due to chance.

Prestige is trickier because then you're asking what kind of sample you want for the kinds of claims you're making. It depends on whether you're making claims about readers, or claims about change. You might have a theory that widely read works are more likely to produce change—but you can also test those theories. I don't know if there's an a priori reason to always base arguments about change on prestigious works. You probably need to test how strong the evidence is that prestigious works are more influential.

We do have some evidence that works ahead of their time line up with certain kinds of prestige signals, and that different kinds of prestige matter to different degrees. For example, you can ask whether being widely reviewed or being perceived as avant-garde correlates most with being ahead of one's time textually.

2.6 Literacy and Practice in the LLM Era

(12) *How has the arrival of ChatGPT and similar systems altered your approach to teaching computational methods, literary history, or theory? What kinds of conceptual literacy do students now need—about training data, evaluation, uncertainty, and institutional incentives—to work responsibly with LLMs in humanities contexts?*

None of us are experts on this—we've had about three years to think about it. My sense is students can do things more quickly. They can learn more quickly and teach themselves things that would have taken a long time before.

But there are pitfalls. When you teach yourself something fast, you may not know the limitations or the ways you can fool yourself with a method. That's true for faculty as well as students. So caution is required.

We also need reflection on the value of getting multiple points of view. From a language model you'll get one point of view—"this method is likely to work"—but it may have blind spots. So you need to consult humans too, because right now language models are liable to share similar blind spots. You want to hear from someone who knows the pitfalls.

What's your experience? Do you find students are using this to teach themselves things?

(13) *Yes. I have them use it and submit chat logs. I expect the final result to be better than just using an LLM. I teach Shakespeare editing, and I wrote an article disputing a recent scholarly agreement about a play's textual status. That position isn't in the model's data, so the model tends to argue against*

mine, reverting to older consensus because that's what it has seen.

That's a great example of the risk: consensus becomes homogenized and becomes hard to change.

(14) *Students add more evidence—newer articles—into the context window and try to force the model to think differently. The result is interesting.*

Are they able to do that—get it to think differently?

(15) *Yes.*

Then in the long run we need some way of getting more difference of opinion out of language models than we're getting right now. It's partly about changing models and partly about teaching humans to be wary of blind spots and hasty consensus.

Students are learning faster, but we have to worry about blind spots. The first answer they get is often a "B" answer; to get an "A," they have to add something else—more input, editing, diversifying the perspectives.

Ultimately we may need a protocol—a skill—to teach people how to search for alternative perspectives: what is the space of opinion here? I don't want just one answer; I want a map of potential perspectives, including what we don't know.

2.7 Anachronism, Authorship, and Theory

(16) *In your work on whether language models can represent the past without anachronism, you argue the problem isn't merely technical. What makes anachronism such a hard challenge for LLM-based humanities research? Do fine-tuning and historical corpus restriction meaningfully solve it, or do they only relocate the epistemic risk?*

There is a technical dimension to this—related to perspectival homogeneity or what you might call mode collapse: a single point of view dominated by training data. With anachronism, the technical problem is that models can't forget. If they have seen the present, they can't forget it exists. That might have technical fixes. Thinking about it as mode collapse also suggests there may be technical fixes—ways to ensure the model looks outside the most common answer.

But the part that is more than technical is hermeneutic. There's always a circularity: our perspective is shaped by what we're trying to understand, which means we can't be totally separate from it in the way some theories of objectivity would require. That problem won't go away. It won't have a technical fix, because it's fundamental. Instead, we need to reflect on it.

(17) *In "The Empirical Triumph of Theory", you suggest LLMs dramatize twentieth-century theories of meaning, and you discuss how instruction tuning reconstructs something like a provisional author-function. What does that imply for how we talk about agency, intention, and interpretation in DH now?*

If the next phase of DH is defined by interpretation and governance, what norms of evidence, disclosure, and evaluation should journals, labs, and graduate training adopt first?

These are very big questions. The challenge of large language models is bigger than DH. Our ideas about human agency and intelligence have rested on subject-centered models that were challenged by structuralism and post-structuralism, and LLMs are challenging them again. That's something we in digital humanities need to think about, but it's part of a much larger social challenge—what happens when people assimilate the insight that language speaks through them. Many people find that scary.

On norms of evidence, disclosure, and evaluation: I'm in favor of open science—sharing code and data—but it gets more complicated with AI. Scholarly institutions are trying right now to reason about what acknowledgments are appropriate for work with AI. I don't have a single right answer. That's something we're hashing out collectively.

(18) *I teach "The Empirical Triumph of Theory" almost every class, because it helps students think about subjectivity and authorship with AI. Even before LLMs, Barthes and Foucault mattered, and the opening paragraph of your article is illuminating.*

I hope language models help those fundamental modes of thinking sink in for students—where they may have slid off in the 1970s as "fancy French words." Now it feels concrete.

(19) *Do you have anything you wanted to add?*

No—those are such thorough questions. Something that strikes me about good interviews is you discover you've forgotten so much of what you've done. Someone who has read through your work may actually know it better than you do, because our memories don't hold it; it's in papers.

That's what's interesting about the interview process: you discover the limits of individual subjectivity—individual consciousness—and how much writing exceeds your brain. Over time you do many different things and can't remember them all; you're a different person now. That's what's fun about this kind of interview: "Oh, I did have that thought."