실행 가이드라인. Comput. Speech Lang., 67:101151.

Leiter, Christoph 및 Steffen Eger. 2024. Prexme! 대규모 프롬프트 탐색을 통한 오픈 소스 대형 언어 모델의 기계 번역과 요약 평가에 대한 학습. In EMNLP, 페이지 11481–11506, 계산 언어학 협회.

Leiter, Christoph, Juri Opitz, Daniel Deutsch, Yang Gao, Rotem Dror, 및 Steffen Eger. 2023. 대형 언어 모델을 설명 가능한 평가 지표로 활용하는 eval4nlp 2023 공동 과제. CoRR, abs/2310.19792.

Li, Junlong, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, 및 Pengfei Liu. 2023a. 정렬 평가를 위한 생성적 판단자. CoRR, abs/2310.05470.

Li, Qintong, Leyang Cui, Lingpeng Kong, 및 Wei Bi. 2023b. 협력 평가: 대형 언어 모델과 인간의 시너지를 탐구하여 개방형 생성 평가. CoRR, abs/2310.19740.

Li, Ruosen, Teerth Patel, 및 Xinya Du. 2023. PRD: 동료 순위 및 논의가 대형 언어 모델 기반 평가를 개선합니다. CoRR, abs/2307.02762.

Li, Yujia, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, 및 Oriol Vinyals. 2022. 경진대회 수준의 코드 생성을 위한 alphacode. Science, 378(6624):1092–1097.

Li, Zongjie, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, 그리고 Yang Liu. 2023c. 분할 및 병합: 대형 언어 모델 기반 평가자의 위치 편향을 정렬하기. CoRR, abs/2310.01432.

Lin, Yen-Ting과 Yun-Nung Chen. 2023. Llm-eval: 대형 언어 모델을 통한 개방형 대화의 통합 다차원 자동 평가. In NLP4ConvAI 2023.

Liu, Yang, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, 및 Chenguang Zhu. 2023a. G-eval: GPT-4를 사용한 자연어 생성 평가 및 개선된 인간 정렬. In EMNLP. Liu, Yixin, Alexander R. Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, 및 Arman Cohan. 2023b. 지시 제어 가능한 요약을 위한 대형 언어 모델의 생성 및 평가 능력 벤치마킹. CoRR, abs/2311.09184.

Liu, Yixin, Alexander R. Fabbri, Pengfei Liu, Dragomir Radev, 및 Arman Cohan. 2023c. 대형 언어 모델을 참조하는 요약 학습에 대하여. CoRR, abs/2305.14239.

Liu, Yongkang, Shi Feng, Daling Wang, Yifei Zhang, 및 Hinrich Schütze. 2023d. 평가할 수 없는 것을 평가하라: 평가할 수 없는 생성된 응답 품질. CoRR, abs/2305.14658.

Liu, Yuxuan, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, 및 Qi Zhang. 2023e. LLM 기반 평가기 보정. CoRR, abs/2309.13308.

Liusie, Adian, Potsawee Manakul, 및 Mark J. F. Gales. 2023. LLM 비교 평가: 대형 언어 모델을 사용한 쌍대 비교를 통한 제로샷 NLG 평가. Computing Research Repository, arxiv:2307.07889.

Lu, Qingyu, Baopu Qiu, Liang Ding, Liping Xie, 및 Dacheng Tao. 2023. 오류 분석 프롬프트가 대형 언어 모델에서 인간과 유사한 번역 평가를 가능하게 한다: ChatGPT에 대한 사례 연구. CoRR, abs/2303.13809.

Luo, Zheheng, Qianqian Xie, 및 Sophia Ananiadou. 2023. Chatgpt를 추상 텍스트 요약을 위한 사실 불일치 평가자로 사용하기. CoRR, abs/2303.15621.

Manakul, Potsawee, Adian Liusie, 그리고 Mark J. F. Gales. 2023. Selfcheckgpt: 생성 대형 언어 모델을 위한 제로 리소스 블랙박스 환각 탐지. In EMNLP.

Mendonça, John, Patrícia Pereira, João Paulo Carvalho, Alon Lavie, 그리고 Isabel Trancoso. 2023. 간단한 LLM 프롬프트가 강력하고 다국어 대화 평가에 최첨단이다. In Proceedings of The Eleventh Dialog System Technology Challenge.

Menick, Jacob, Maja Trebacz, Vladimir Mikulik, John Aslanides, H. Francis Song, Martin J. Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, 그리고 Nat McAleese. 2022. 언어 모델이 검증된 인용으로 답변을 지원하도록 가르치기. CoRR, abs/2203.11147.

Murugadoss, Bhuvanashree, Christian Poelitz, Ian Drosos, Vu Le, Nick McKenna,

Carina Suzana Negreanu, Chris Parnin, and Advait Sarkar. 2024. 평가자를 평가하기: 대형 언어 모델의
작업 평가 지침 준수 측정. arXiv 사전 인쇄 arXiv:2408.08781.

Naismith, Ben, Phoebe Mulcaire, and Jill Burstein. 2023. GPT-4를 사용한
작성된 담화 일관성의 자동 평가. BEA@ACL.

Nakano, Reiichiro, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. WebGPT: 인간 피드백을 통한
브라우저 지원 질문 응답. CoRR, abs/2112.09332.

Ni'mah, Iftitahu, Meng Fang, Vlado Menkovski, and Mykola Pechenizkiy. 2023. NLG 평가 지표의 상관 분석을 넘어서: 경험적 지표
선호 체크리스트. In ACL (1), 페이지
1240–1266, 계산 언어학 협회.

Ostyakova, Lidiia, Veronika Smilga, Kseniia Petukhova, Maria Molchanova, and Daniel Kornev. 2023. ChatGPT 대
크라우드소싱 대 전문가: 발화 기능으로 열린
도메인 대화를 주석 처리하기. In 제24회 연례
전문 위원회 회의록
담화 및 대화, 페이지 242–254,
계산 언어학 협회, 프라하, 체코.
Linguistics, Prague, Czechia.

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulma n, Jacob Hilton, Fraser Kelton, Luke Miller, Maddi e Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, 그리고 Ryan Lowe. 2022. 인
간 피드백을 통한 지침을 따르는 언어 모델 훈련. In Neur IPS.

Papineni, Kishore, Salim Roukos, Todd Ward, 및 Wei-Jing Zhu. 2002. BLEU: 기계 번역의 자동 평가 방법. In ACL.

Pozdniakov, Stanislav, Jonathan Brazil,
Solmaz Abdi, Aneesha Bakharia, Shazia Sadiq, Dragan Ga□evi□, Mohsen Mirhoseini, Catherine Lynch, Nicholas King, Harsha Nori, 및 Saleema Amershi.
2023. LLM을 활용한 감사에서 인간-AI 협업 지원. AIES.

Ribeiro, Marco Túlio 및 Scott M. Lundberg.
2022. NLP 모델의 적응형 테스트 및 디버깅. In ACL (1).

Saha, Swarnadeep, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, 및 Xian L
2022. Adaptive testing and debugging of NLP models. In *ACL (1)*.

Sheth, Sreyas Mohan, 및 Mitesh M Khapra. 2021. NLG 평가 지표 평가를 위한 섭동 체크리스
Empirical Methods in Natural LanguageProcessing,
페이지 7219–7234.

Sai, Ananya B., Akash Kumar Mohankumar, 및 Mitesh M. Khapra. 2023. NLG 시스템에 사
Saunders, William, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, 및 Ja n Leike. 2022. 인간 평가자를 지원하는 자기 비판 모델. CoRR, abs/2206.05802.

Shankar, Shreya, J.D. Zamfirescu-Pereira, *Processing*, pages 7219–7234.

Sai, Ananya B., Akash Kumar Mohankumar, and Mitesh M. Khapra. 2023. A survey of evaluation metrics used for NLG systems. *ACM Comput. Surv.*, 55(2):26:1–26:39.

Saunders, William, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. *CoRR*, abs/2206.05802.

Shankar, Shreya, J.D. Zamfirescu-Pereira, Bjoern Hartmann, Aditya Parameswaran, 그리고 Ian Arawjo. 2024. 누가 평가자를 검증하는가? 대형 언어 모델의 출력에 대한 인간의 선호와 일치하는 LLM 지원 평가. 제37회 회의록에서
연례 ACM 사용자 인터페이스
소프트웨어 및 기술, UIST '24,
전산 기계 협회,
뉴욕, NY, 미국.

Shen, Chenhui, Liying Cheng, Xuan-Phi Nguyen, Yang You, 그리고 Lidong Bing. 2023. 대형 언어 모델은 아직
추상적 요약을 위한 인간 수준 평가자가 아니다. EMNL P (발표자료).

Sheng, Shuqian, Yi Xu, Tianhang Zhang, Zanwei Shen, Luoyi Fu, Jiaxin Ding, Lei Zhou, Xiaoying Gan, Xinbing Wang, 및 Chenghu Zhou. 2024. Repeval: LLM 표현을 통한
자동 평가 방법. In ACL. EMNLP, 페이지 7019–7033, 전산 언어학 협회.

Shu, Lei, Nevan Wichers, Liangchen Luo, Yun Zhu, Yinxiao Liu, Jindong Chen, 2024. 대형 언어 모델이 사용자 인터페이스를 만나다: 피
Lei Meng. 2023. Fusion-eval: LLM과 평가자 통합. Co RR,
abs/2311.09204.

Stent, Amanda, Matthew Marge, 및 Mohit Singhai. 2005. 변이가 존재하는 경우의
생성을 위한 평가 방법 평가. CICLing, 컴퓨터 과학 강의 노트의
3406 권, 페이지

계산 언어학 방금 수락된 MS.
https://doi.org/10.1162/coli_a_00561

341–351, 스프링거.
Sun, Tianxiang, Junliang He, Xipeng Qiu, 그리고 Xuanjing Huang. 2022. Bertscore는 부정확하다: 텍스트 생성에 대한 언어 모델 기반 평가 지표의 사회적 편향. In EMNLP.
Törnberg, Petter. 2023. Chatgpt-4 전문가와 군중 작업자보다 정치적 트위터 메시지를 배정하는 데 제로샷 학습으로 우수함. CoRR, abs/2304.06588.
Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, 및 Thomas Scialom. 2023. Llama 2: 오픈 기초 및 세부 조정된 채팅 모델. CoRR, abs/2307.09288.
Varshney, Neeraj, Wenlin Yao, Hongming Zhang, Jianshu Chen, 및 Dong Yu. 2023. 적기에 시간을 아낀다: 낮은 신뢰도를 가진 생성을 검증하여 LLM의 환각을 탐지하고 완화하기. CoRR, abs/2307.03987.
왕, 지안, 윤롱 리앙, 판동 멍. 하오샹 시, 지쉬 리, 진안 쉬, 잔펑 퀴, 그리고 지에 저우. 2023a. ChatGPT는 좋은 자연어 생성(NLG) 평가자일까? 초기 연구. In 제4회 새로운 프론티어에서의 요약 워크숍.
왕, 페이이, 레이 리, 리앙 첸, 다웨이 저우, 빙화이 린, 윤보 카오, 귀 리우, 톈위 리우, 그리고 즈팡 수이. 2023b. 대형 언어 모델은 공정한 평가자가 아니다. CoRR, abs/2305.17926.
왕, 티안루, 일리아 쿨리코프, 올가 골로브네바, 핑 유, 웨이제 유안, 제인 드위베디-유, 리차드 유안제 팡, 마리암. 파젤-자란디, 제이슨 웨스턴, 그리고 시안 리. 2024. 자기 학습된 평가자. arXiv preprint arXiv:2408.02666.
왕, 티안루, 핑 유, 샤오칭 엘렌 탄, 션 오브라이언, 라마칸트 파순루, 제인 드위베디-유, 올가 골로브네바, 루크. 젯틀모이어, 마리암 파젤-자란디, 그리고 아슬리 첼리키요르마즈. 2023c. Shepherd: 언어 모델 생성을 위한 비평자. CoRR, abs/2308.04592.
왕, 야칭, 지에푸 장, 밍양 장, 청 리, 이 리앙, 취아오주 메이, 그리고 마이클 벤더스키. 2023d. 대형 언어 모델을 사용한 개인화된 텍스트 생성의 자동 평가. CoRR, abs/2310.11593.
왕, 이둥, 주오하오 유, 정란 젠, 린이 양, 쿤샹 왕, 하오 첸, 차오야 장, 루이 시에, 진돈 왕, 싱 시에, 웨이 예, 시쿤 장, 그리고 위에 장. 2023e. Pandalm: LLM 지시 조정 최적화를 위한 자동 평가 벤치마크. CoRR, abs/2306.05087.
왕, 지판, 고타로 후나코시, 그리고 마나부 오쿠무라. 2023. 질문 생성을 위한 자동 답변 가능성 평가. CoRR, abs/2309.12546.
와이즈먼, 샘, 스튜어트 M. 시버, 그리고 알렉산더 M. 러시. 2017. 데이터-투-문서 생성의 도전 과제. In EMNLP, 페이지 2253–2263, 계산 언어학 협회. Computational Linguistics.
우, 밍하오와 알함 피크리 아지. 2023. 형식에 대한 내용: 대형 언어 모델에 대한 평가 편향. CoRR, abs/2307.03025.
우, 닝, 밍 공, 링전 쇼우, 샤이닝. Liang과 Daxin Jiang. 2023. 대형 언어 모델은 요약 평가를 위한 다양한 역할을 합니다. NLPCC (1).
Xiao, Ziang, Susu Zhang, Vivian Lai, 그리고 Q. Vera Liao. 2023. 평가 지표 평가: 측정 이론을 사용한 자연어 생성(NLG) 평가 지표 분석을 위한 프레임워크. EMNLP, 페이지 10967–10982, 컴퓨터 언어학 협회. Linguistics.
Xie, Zhuohan, Miao Li, Trevor Cohn, 그리고 Jey Han Lau. 2023. Deltascore: 세부적인 이야기 평가를 위한 섭동. EMNLP (Findings).
Xu, Wenda, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, 그리고 Lei Li. 2023.

계산 언어학 방금 수락된 MS.
https://doi.org/10.1162/coli_a_00561

INSTRUCTSCORE: 설명 가능한
텍스트 생성 평가를 위한 자동화된
피드백 방향으로. EMNLP.
Ye, Seonghyeon, Doyoung Kim, Sungdong
Kim, Hyeonbin Hwang, Seungone Kim,
Yongrae Jo, James Thorne, Juho Kim, 및
Minjoon Seo. 2023a. FLASK: 정밀한
언어 모델 평가
정렬 기술 집합에 기반하여. CoRR,
abs/2307.10928.
Ye, Xi, Srinivasan Iyer, Asli Celikyilmaz,
Veselin Stoyanov, Greg Durrett, 및
Ramakanth Pasunuru. 2023b.
효과적인
맥락 내 학습을 위한 보완 설명. ACL (발견).
Yin, Kayo 및 Graham Neubig. 2022.
대조적인 설명으로
언어 모델 해석. EMNLP.
Yuan, Peiwen, Shaoxiong Feng, Yiwei Li,
Xinglin Wang, Boyuan Pan, Heda Wang,
및 Kan Li. 2024. Batcheval: 인간과 유사한
텍스트 평가를 지향하다. CoRR,
abs/2401.00437.
Yuan, Weizhe, Graham Neubig, 및 Pengfei
Liu. 2021. Bartscore: 생성된
텍스트를 텍스트 생성으로 평가. NeurIPS.
Zhang, Chen, Luis Fernando D'Haro, Yiming
Chen, Malu Zhang, 및 Haizhou Li.
2023a. 대형 언어 모델의
효과성에 대한 포괄적 분석으로
자동 대화 평가자로서. CoRR,
abs/2312.15407.
Zhang, Susan, Stephen Roller, Naman Goyal,
Mikel Artetxe, Moya Chen, Shuohui Chen,
Christopher Dewan, Mona T. Diab, Xian
Li, Xi Victoria Lin, Todor Mihaylov, Myle
Ott, Sam Shleifer, Kurt Shuster, Daniel
Simig, Punit Singh Koura, Anjali Sridhar,
Tianlu Wang, 및 Luke Zettlemoyer. 2022.
OPT: 공개 사전 훈련된 트랜스포머
언어 모델. CoRR, abs/2205.01068.
Zhang, Xinghua, Bowen Yu, Haiyang Yu,
Yangyu Lv, Tingwen Liu, Fei Huang,
홍보 수와 용빈 리. 2023b. 더 넓고 깊은 대형 언어 모델 네
트워크는 더 공정한 대형 언어 모델 평가자이다. CoRR, ab
s/2308.01862.
장, 양준, 펑지에 렌, 마르텐
드 라이크. 2021. 대화에서 악의성을 평가하기 위한 인간-

기계 협력 프레임워크. ACL/IJCNLP (1).

장, 잉, 슈테판 보겔, 알렉스
와이블. 2004. BLEU/NIST 점수 해석: 더 나은 시스템
을 갖기 위해 얼마나 개선해야 할까? LREC, 유럽 언어
자원

협회.
정, 리안민, 웨이린 장, 잉
셩, 시위안 주앙, 장하오 우,
용하오 주앙, 지 린, 주하오 리,
다쳉 리, 에릭 P. 싱, 하오 장,
조셉 E. 곤잘레스, 그리고 이온 스토이카. 2023.
대형 언어 모델을 평가자로 사용하여 mt-bench와 대화형
봇 아레나를 평가한다. CoRR, abs/2306.05685.
조우, 케이틀린, 수 린 블로젯, 아담
트리슐러, 할 다우메 3세, 카히르 술레만,

알렉산드라 올테아누. 2022.
자연어 생성 평가 분석:

평가 관행, 가정 및 그들의 의미. In NAACL-HLT, 페이지 314–324, 계산 언어학 협회.
저우, 용신, 파비앵 링바르, 그리고

프랑수아 포르테. 2023. 생성된 의료 텍스트 보고서 평가 방법 조사. ClinicalNLP@AC
주, 량후이, 싱강 왕, 그리고 신롱
왕. 2023. Judgelm: 미세 조정된 대형 언어 모델은 확장
가능한 평가자이다.
CoRR, abs/2310.17631.
주앙, 지유, 치광 천, 롱쉬안 마,
밍다 리, 이 한, 유산 치안, 하오펑

바이, 지시안 펑, 웨이난 장, 그리고 팅
리우. 2023. 핵심 역량의 렌즈를 통해: 대형 언어 모델 평가
조사. CoRR, abs/2308.07902.