실무 지침. Comput. Speech Lang., 67:101151.

Leiter, Christoph와 Steffen Eger. 2024. Prexme! 기계 번역과 요약 평가를 위한 오픈소스 대형 언어 모델의 대규모 프롬프트 탐색. EMNLP, 11481-11506 페이지, Association for Computational Linguistics.

Leiter, Christoph, Juri Opitz, Daniel Deutsch, Yang Gao, Rotem Dror, 그리고 Steffen Eger. 2023. eval4nlp 2023 설명 가능한 지표로서의 대형 언어 모델 프롬프팅에 관한 공유 과제. CoRR, abs/2310.19792.

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

Li, Qintong, Leyang Cui, Lingpeng Kong, 그리고 Wei Bi. 2023b. 협업 평가: 개방형 생성 평가를 위한 대형 언어 모델과 인간의 시너지 탐구. CoRR, abs/2310.19740.

Li, Ruosen, Teerth Patel, 그리고 Xinya Du. 2023. PRD: 동료 순위 및 토론을 통한 대형 언어 모델 기반 평가 개선.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, 그리고 Oriol Vinyals. 2022. AlphaCode를 이용한 대회 수준의 코드 생성. Science, 378(6624):1092–1097.

Li, Zongjie, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, Yang Liu. 2023c. 대형 언어 모델 기반 평가자들의 위치 편향 분리 및 병합. CoRR, abs/2310.01432.

Lin, Yen-Ting 및 Yun-Nung Chen. 2023. Llm-eval: 대형 언어 모델을 활용한 개방형 대화를 위한 통합된 다차원 자동 평가. In NLP4ConvAI 2023.

Liu, Yang, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, Chenguang Zhu. 2023a. G-eval: 인간 정렬이 개선된 GPT-4를 활용한 자연어 생성 평가. In EMNLP.

Liu, Yixin, Alexander R. Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, 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, Arman Cohan. 2023c. 대형 언어 모델을 참조로 활용한 요약 학습에 관하여. CoRR, abs/2305.14239.

Liu, Yongkang, Shi Feng, Daling Wang, Yifei Zhang, 및 Hinrich Schütze. 2023d. 평가할 수 없는 것을 평가하기: 평가 불가능한 생성된 응답의 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. 대형 언어 모델 기반 평가자의 보정. CoRR, abs/2309.13308.

Liusie, Adian, Potsawee Manakul, 및 Mark J. F. Gales. 2023. 대형 언어 모델의 비교 평가: 대형 언어 모델을 사용한 쌍별 비교를 통한 제로샷 자연어 생성 평가. Computing Research Repository, arxiv:2307.07889.

Lu, Qingyu, Baopu Qiu, Liang Ding, Liping Xie, 및 Dacheng Tao. 2023. 오류 분석 프롬프팅을 통한 대형 언어 모델의 인간다운 번역 평가 실현: ChatGPT 사례 연구.CoRR, abs/2303.13809.

Luo, Zheheng, Qianqian Xie, 및 Sophia Ananiadou. 2023. ChatGPT를 이용한 추상적 텍스트 요약의 사실적 비일관성 평가. CoRR, abs/2303.15621.

Manakul, Potsawee, Adian Liusie, 그리고 Mark J. F. Gales. 2023. SelfCheckGPT: 생성형 대형 언어 모델을 위한 무자원 블랙박스 환각 탐지. In EMNLP.

Mendonça, John, Patrícia Pereira, João Paulo Carvalho, Alon Lavie, 그리고 Isabel Trancoso. 2023. 간단한 LLM 프롬프팅이 견고하고 다국어 대화 평가의 최신 기술이다. 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. 검증된 인용구로 답변을 뒷받침하도록 언어 모델 교육하기. CoRR, abs/2203.11147.

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

Carina Suzana Negreanu, Chris Parnin, 그리고 Advait Sarkar. 2024. 평가자 평가하기: 과제 평가 지침에 대한 대형 언어 모델의 준수도 측정. arXiv 사전인쇄 arXiv:2408.08781.

Naismith, Ben, Phoebe Mulcaire, 그리고 Jill Burstein. 2023. GPT-4를 사용한 문장 담화 일관성의 자동화된 평가. In 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, 그리고 John Schulman. 2021. Webgpt: 인간 피드백을 통한 브라우저 지원 질의응답. CoRR, abs/2112.09332.

Ni'mah, Iftitahu, Meng Fang, Vlado Menkovski, and Mykola Pechenizkiy. 2023. 상관관계 분석을 넘어선 자연어 생성 평가 지표: 경험적 지표 선호도 체크리스트. In ACL (1), pages 1240–1266, Association for Computational Linguistics.

Ostyakova, Lidiia, Veronika Smilga, Kseniia Petukhova, Maria Molchanova, and Daniel Kornev. 2023. ChatGPT vs. 크라우드소싱 vs. 전문가: 개방형 대화에 대한 발화 기능 주석 달기. In Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 242–254, Association for Computational 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 Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, 그리고 Ryan Lowe. 2022. 인간 피드백을 통한 지시사항 준수를 위한 언어 모델 훈련. In NeurIPS.

Papineni, Kishore, Salim Roukos, Todd Ward, 그리고 Wei-Jing Zhu. 2002. Bleu: 기계 번역의 자동 평가를 위한 방법. In ACL.

Pozdniakov, Stanislav, Jonathan Brazil, Solmaz Abdi, Aneesha Bakharia, Shazia Sadiq, Dragan Ga□evi□, Paul Denny, 그리고 Hassan Khosravi. 2024. 대형 언어 모델과 사용자 인터페이스의 만남: 피드백 제공의 사례. Computers and Education: Artificial Intelligence, 7:100289.

Rastogi, Charvi, Marco Túlio Ribeiro, Nicholas King, Harsha Nori, 그리고 Saleema Amershi. 2023. LLM을 이용한 LLM 감사에서의 인간-AI 협업 지원. In AIES.

Ribeiro, Marco Túlio 그리고 Scott M. Lundberg. Bjoern Hartmann, Aditya Parameswaran, 그리고 Ian Arawjo. 2024. 검증자를 누가 검증하는가? 인간 선호도와 LLM 출력의 LLM 보조 평가 정렬. 제37회 연례 ACM 사용자 인터페이스 소프트웨어 및 기술 심포지엄, UIST '24, Association for Computing Machinery, New York, NY, USA.

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

Sheng, Shuqian, Yi Xu, Tianhang Zhang, Zanwei Shen, Luoyi Fu, Jiaxin Ding, Lei Zhou, Xiaoying Gan, Xinbing Wang, 그리고 Chenghu Zhou. 2024. Repeval: LLM 표현을 통한 효과적인 텍스트 평가. EMNLP, pages 7019-7033, Association for Computational Linguistics.

Shu, Lei, Nevan Wichers, Liangchen Luo, Yun Zhu, Yinxiao Liu, Jindong Chen, 그리고 Lei Meng. 2023. Fusion-eval: LLM과 평가자의 통합. CoRR, abs/2311.09204.

Stent, Amanda, Matthew Marge, 그리고 Mohit Singhai. 2005. 변이가 존재하는 상황에서의 생성을 위한 평가 방법 평가. CICLing, 제3406권 Lecture Notes in Computer Science, 페이지

341-351, Springer.
Sun, Tianxiang, Junliang He, Xipeng Qiu, 및 Xuanjing Huang. 2022. BERTScore는 불공정하다: 텍스트 생성을 위한 언어 모델 기반 지표의 사회적 편향에 관하여. EMNLP에서.
Törnberg, Petter. 2023. ChatGPT-4 제로샷 학습으로 정치적 트위터 메시지를 주석하는 데 있어 전문가와 군중 작업자보다 더 우수한 성능을 보임.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, Zettlemoyer, Maryam Fazel-Zarandi, 및 Asli Celikyilmaz. 2023c. Shepherd: 언어 모델 생성을 위한 비평가. CoRR, abs/2308.04592.
Wang, Yaqing, Jiepu Jiang, Mingyang Zhang, Cheng Li, Yi Liang, Qiaozhu Mei, 및 Michael Bendersky. 2023d. 대형 언어 모델을 활용한 개인화된 텍스트 생성의 자동화된 평가. CoRR, abs/2310.11593.
Wang, Yidong, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, 및 Yue Zhang. 2023e. PandaLM: LLM 지시어 튜닝 최적화를 위한 자동 평가 벤치마크. CoRR, abs/2306.05087.
Wang, Zifan, Kotaro Funakoshi, 및 Manabu Okumura. 2023. 질문 생성을 위한 자동 답변 가능성 평가. CoRR, abs/2309.12546.
Wiseman, Sam, Stuart M. Shieber, 및 Alexander M. Rush. 2017. 데이터-문서 생성의 도전 과제. In EMNLP, pages 2253-2263, Association for Computational Linguistics.
Wu, Minghao와 Alham Fikri Aji. 2023. 대형 언어 모델에 대한 평가 편향에서의 형식과 내용. CoRR, abs/2307.03025.
Wu, Ning, Ming Gong, Linjun Shou, Shining Liang, 그리고 Daxin Jiang. 2023. 요약 평가를 위한 다양한 역할 수행자로서의 대형 언어 모델. NLPCC (1)에서.
Xiao, Ziang, Susu Zhang, Vivian Lai, 그리고 Q. Vera Liao. 2023. 평가 지표 평가하기: 측정 이론을 사용한 자연어 생성 평가 지표 분석을 위한 프레임워크. EMNLP, 페이지 10967-10982, Association for Computational Linguistics.
Xie, Zhuohan, Miao Li, Trevor Cohn, 그리고 Jey Han Lau. 2023. Deltascore: 교란을 통한 세밀한 스토리 평가. EMNLP (Findings)에서. Xu, Wenda, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, 그리고 Lei Li. 2023.

전산 언어학 논문 게재 확정.
https://doi.org/10.1162/coli_a_00561

INSTRUCTSCORE: 자동 피드백을 통한 설명 가능한 텍스트 생성 평가를 향하여. In EMNLP.

Ye, Seonghyeon, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, 그리고 Minjoon Seo. 2023a. FLASK: 정렬 기술 세트에 기반한 세분화된 언어 모델 평가. CoRR, abs/2307.10928.

Ye, Xi, Srinivasan Iyer, Asli Celikyilmaz, Veselin Stoyanov, Greg Durrett, 그리고 Ramakanth Pasunuru. 2023b. 효과적인 문맥 내 학습을 위한 보완적 설명. In ACL (Findings).

Yin, Kayo 그리고 Graham Neubig. 2022. 대조적 설명을 통한 언어 모델 해석. In EMNLP.

Yuan, Peiwen, Shaoxiong Feng, Yiwei Li, Xinglin Wang, Boyuan Pan, Heda Wang, 그리고 Kan Li. 2024. Batcheval: 인간과 유사한 텍스트 평가를 향하여. CoRR, abs/2401.00437.

Yuan, Weizhe, Graham Neubig, 그리고 Pengfei Liu. 2021. Bartscore: 생성된 텍스트를 텍스트 생성으로 평가하기. In NeurIPS.

Zhang, Chen, Luis Fernando D'Haro, Yiming Chen, Malu Zhang, 그리고 Haizhou Li. 2023a. 자동 대화 평가자로서의 대형 언어 모델 효과성에 대한 포괄적 분석. 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: 개방형 사전 학습 변환기 언어 모델. CoRR, abs/2205.01068.

Zhang, Xinghua, Bowen Yu, Haiyang Yu, Yangyu Lv, Tingwen Liu, Fei Huang, Hongbo Xu, 및 Yongbin Li. 2023b. 더 넓고 더 깊은 대형 언어 모델 네트워크가 더 공정한 대형 언어 모델 평가자이다. CoRR, abs/2308.01862.

Zhang, Yangjun, Pengjie Ren, 및 Maarten de Rijke. 2021. 대화의 악의성을 평가하기 위한 인간-기계 협업 프레임워크. In ACL/IJCNLP (1).

Zhang, Ying, Stephan Vogel, 및 Alex Waibel. 2004. BLEU/NIST 점수 해석: 더 나은 시스템을 위해서는 얼마나 많은 개선이 필요한가? In LREC, European Language Resources Association.

Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, 및 Ion Stoica. 2023.

Zouhar, Vilém, Tom Kocmi, 그리고 Mrinmaya Sachan. 2025. AI 지원 인간의 기계 번역 평가.