피드백 기법을 이용한 LLama2 모델 기반의 Zero-Shot 문서 그라운딩된 대화 시스템 성능 개선

정민교°[1], 홍범석[2], 최원석[3], 한영섭[4], 전병기[5], 나승훈[6]
전북대학교[1,6], LG 유플러스[2,3,4,5]
{jungmk, nash}@jbnu.ac.kr, {bshong, wseokchoi, yshan042, bkjeon}@lguplus.co.kr

LLaMA2 Models with Feedback for Improving Document-Grounded Dialogue System

Min-Kyo Jung°[1], Beomseok Hong[2], Wonseok Choi[3], Youngsub Han[4], Byoung-Ki Jeon[5], Seung-Hoon Na[6]
Jeonbuk National University[1,6], LG Uplus[2,3,4,5]

요 약

문서 그라운딩된 대화 시스템의 응답 성능 개선을 위한 방법론을 제안한다. 사전 학습된 거대 언어 모델 LLM(Large Language Model)인 Llama2 모델에 Zero-Shot In-Context learning을 적용하여 대화 마지막 유저 질문에 대한 응답을 생성하는 태스크를 수행하였다. 본 연구에서 제안한 응답 생성은 검색된 top-1 문서와 대화 기록을 참조해 초기 응답을 생성하고, 생성된 초기 응답을 기반으로 검색된 문서를 대상으로 재순위화를 수행한다. 이 후, 특정 순위의 상위 문서들을 이용해 최종 응답을 생성하는 과정으로 이루어진다. 검색된 상위 문서를 이용하는 응답 생성 방식을 Baseline으로 하여 본 연구에서 제안한 방식과 비교하였다. 그 결과, 본 연구에서 제안한 방식이 검색된 결과에 기반한 실험에서 Baseline 보다 F1, Bleu, Rouge, Meteor Score가 향상한 것을 확인 하였다.

주제어: Llama2, Large Language Model, In-Context learning

1. 서론

대화 시스템이란 사용자 질의에 대해 에이전트가 적절한 응답을 생성하는 시스템을 의미한다. 에이전트는 사용자 질의에 대한 적절한 응답 생성과 자연스러운 대화 흐름을 위해 마지막 질의 전 사용자와 에이전트 간의 대화 기록을 참조한다. 이러한 방식의 대화 시스템은 모델 내부에 학습된 지식을 이용해 사용자 질의에 대한 응답을 생성한다. 모델이 많은 데이터를 이용해 충분히 학습이 되어있다면 사용자 질의에 대한 적절한 응답을 생성할 수 있지만 학습하지 못한 질의에 대해서는 적절하지 못한 응답을 생성할 수 있다. 또한 학습되지 않은 도메인의 지식이 필요한 응답을 생성하지 못한다는 단점이 존재한다. 내부 지식에 의존한 대화 시스템을 개선하기 위해 사용자 발화와 연관된 문서를 검색하고, 검색된 문서가 그라운딩된 응답을 생성하는 문서 그라운딩 대화 시스템(Document grounded Dialogue system)이 제안되었다. 본 연구에서는 사용자 질의에 대한 적절한 응답을 생성하기 위해 문서 그라운딩된 대화시스템 태스크를 수행하였다.

문서 그라운딩 대화 시스템 태스크는 검색과 MRC(Machine Reading Comprehension) 두 단계로 수행된다. 검색 단계는 문서 코퍼스 내에서 마지막 사용자 질의 및 대화 기록과 가장 연관된 문서를 순위화 하는 것을 목적으로 수행되고, MRC 단계는 검색된 문서를 대상으로 사용자 질의에 적절한 응답을 생성하는 것을 목적으로 주로 transformer[1] 기반의 BART[2], T5[3]을 이용해 수행된다. 본 연구에서는 MRC 단계의 응답 생성 성능 개선을 위해 피드백 기반의 In-Context learning 방법론을 대화 시스템 태스크에 적용하였다. 또한 MRC 단계에서 참조할 문서를 위해 검색 단계에서는 검색 분야에서 준수한 성능을 보이는 DPR(Dense Passage Retrieval)[4] 모델을 사용하였다.

자연어 처리 분야에서 transformer의 등장은 다양한 태스크의 성능 향상을 가져왔고, transformer의 인코더, 디코더 구조 기반의 언어 모델이 등장하였다. 특히, 디코더를 사용하는 언어 모델은 생성 태스크에 적합해 QA, 요약, 대화 시스템 등에서 우수한 성능을 보였다. 또한 초기 언어 모델이 등장했을 때와 비교해 최근 등장한 수 십, 수 백 Billion 단위의 파라미터를 사전 학습한 LLM은 특정 프롬프트를 입력하는 In-Context learning이 가능해 내부 파라미터를 업데이트하는 fine-tuning의 필요성을 줄일 수 있었다. 이러한 장점을 이용하기 위해 본 연구에서는 대규모 파라미터를 사전학습한 Llama2[5] 모델을 이용해 Zero-Shot In-Context learning을 수행하고, Baseline과 비교해 응답 생성 성능 향상을 보였다.

최근, 방대한 지식을 학습한 LLM의 내부 지식을 이용한 생성 태스크에서 생성 성능의 향상을 위해 피드백 방법론이 사용되고 있다[6, 7]. LLM이 생성한 결과에 대해 피드백을 입력해 입력을 수정하였을 때, 생성 품질의 향상을 보였다. 기존 피드백 방법론을 적용한 LLM 연구들은 성능 향상을 보였지만 인간이 직접 피드백을 작성해야 한다는 문제점이 있다. 이러한 문제점을 해결하기 위해 본 연구에서는 피드백을 자동으로 생성하는 피드백 선별 방법론을 제안하였다. 피드백 선별을 통해 피드백 작성에 사용되는 비용을 줄일 수 있었고, 선별된 피드백을 통해 대화 마지막 사용자 질의에 대한 응답 생성 성능 향상을 보였다.

본 논문의 기여는 다음과 같다.

- 검색 증강형 LLM을 통해 방대한 양의 파라미터를 사전 학습한 LLM 내부 지식과 검색 모듈을 통해 검색된 외부 지식을 함께 사용해 Baseline 대비 검색된 문서 그라운딩된 대화 시스템 태스크 성능 향상을 보였다.
- 프롬프트를 이용하는 In-Context learning을 이용해 fine-tuning 시간 소비를 줄일 수 있었고, 별도의 예시가 없이 학습하는 Zero-Shot learing의 Llama2 성능을 확인하였다.
- 피드백 기법을 사용해 LLM의 초기 응답을 개선하여 사용자 질의에 대한 응답 생성의 성능 향상을 보였고, 피드백을 자동 선별하는 기법을 제안하였다.

2. 관련 연구

2.1 검색 모델

주어진 질의와 연관된 문서를 검색하기 위해 초기에는 희소 벡터 방법론을 사용하였다. 희소 벡터 방법론은 표현해야할 문장 수가 증가할 수록 차원 수가 비례하여 증가한다는 한계가 있어 이를 극복한 밀집 벡터 임베딩 방법론의 검색 모델이 등장하였다. 밀집 벡터 임베딩 방법론에는 질의와 문서를 각각 인코딩하고, 인코딩된 두 결과를 내적하여 유사도를 계산하는 하는 DPR 모델이 있다. 본 연구에서는 이러한 DPR을 이용해 마지막 사용자 질의와 문서 코퍼스내의 유사도를 계산하여 상위 k개의 문서를 검색하고, 응답 생성에 사용하였다.

2.2 Large Language Model

다양한 자연어 처리 태스크의 성능이 대용량의 코퍼스를 사전 학습한 언어 모델의 등장으로 향상되었다. 사전 학습 언어 모델 등장 초기에 BERT[8], RoBerta[9], ELECTRA[10] 등 인코더 기반의 모델들이 분류, 상호참조 해결, 의존 파싱 등 다양한 태스크에서 우수한 성능을 보였다. 이 후, 인코더-디코더 또는 디코더 구조 기반의 T5, BART, GPT 계열[11, 12] 모델 등의 등장으로 여러 생성 관련 태스크에서 우수한 성능을 보였다.

여러 언어 모델들이 확장하면서 동시에 파라미터 사이즈가 수 십, 수 백 Billion 이상으로 방대하게 증가하였다. 파라미터 사이즈의 증가로 언어 모델의 내부 지식의 능력이 향상되었고, 이에 따라 내부 지식을 활용하는 연구가 수행되었다. Llama 계열 모델[13]과 GPT 계열 모델 등과 같이 프로프트를 튜닝하여 학습하는 In-Context learning이 등장하였고, 입력에 예시를 주는 방법에 따라 Zero-Shot, Few-Shot 등의 학습 방법이 등장 하였다. 본 연구에서는 오픈소스로 공개된 Llama2-(7B)를 기반으로 Zero-Shot leaning을 수행하였다.

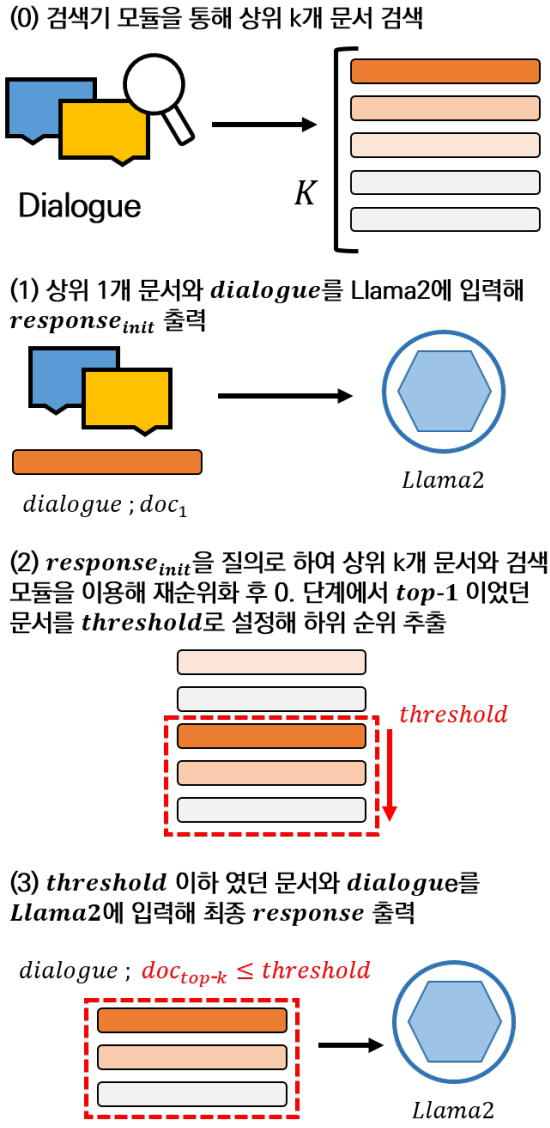

그림 1. In-Context learning 기반 문서 그라운딩된 대화 시스템 모델 개요

2.3 Natural Language Feedback

방대한 규모의 파라미터를 사전 학습한 LLM의 등장으로 다양한 태스크에서 성능이 향상되었지만 존재하지 않는 정보 또는 잘못된 정보를 생성하는 환각 문제가 동시에 등장하였다[14, 15]. 환각 문제를 해결하기 위해 LLM에 의해 생성된 문장에 점수를 부여해 해결하려는 연구가 수행되었다[16, 17]. 점수를 부여하는 방식의 피드백은 사람이 구체적으로 원하는 피드백을 언어 모델에 전달하기 어렵다는 한계가 있다. 이에 따라, LLM에 의해 생성된 문장에 피드백을 구체적으로 전달할 수 있는 자연어 피드백 전달 방법론이 제안되었다. 해당 연구에서 제안된 방법론은 크게 두 단계로 수행된다. 먼저, 초기 생성된 문장에 대해 사람이 피드백을 작성하고, 피드백과 함께 초기 생성된 문장을 LLM에 다시 입력하여 k개의 문장을 다시 생성

Step	Prompt
초기 응답 생성 (1)	please, refer to the document and dialogue to respond to the question.\n\n [document]\n {doc}\n\n [dialogue]\n{history}\n\n[question]\n{last_turn}\n\n[response]\n
최종 응답 생성 (3)	please, refer to the document and dialogue to respond to respond to the question.\n\n[document]\n {doc k}...\n\n[dialogue]\n{history}\n\n[question]\n{last_turn}\n\n[response}\n

표 1. Llama2 입력을 위한 프롬프트 템플릿으로 {doc}는 검색된 문서, {history}는 대화 기록, {last_turn}는 마지막 질의 텍스트가 삽입되는 위치를 의미한다. 또한 {doc k}는 k개의 검색된 문서가 순차적으로 삽입되는 위치를 의미한다. Step의 (1)과 (2)는 그림 1의 (1)단계, (2)단계를 의미한다.

하는 과정을 수행한다. 이 후, 생성된 k개의 문장과 피드백 간의 유사도를 계산하여 상위 유사도를 보인 생성된 문장을 대상으로 fine-tuning을 수행한다. 이는 자연어로 섬세한 피드백을 줄 수 있다는 강점이 있지만 사람이 직접 피드백을 작성해야하는 한계가 존재한다. 이러한 한계를 극복하고자 본 연구에서는 검색된 문서에서 피드백을 선별하고, 선별된 피드백을 LLM에 전달하는 방법을 제안하였다.

3. 제안 방법

본 연구에서 제안하는 문서 그라운딩된 대화 시스템 구조는 그림 1과 같은 과정으로 수행된다. (0)단계와 같이 대화 *dialogue*를 질의로 하여 연관된 문서 상위 k개를 검색기(DPR) *retrieval*를 이용해 검색한다. 이후, (1)단계와 같이 검색된 *top*-1 문서 $doc_{top\text{-}1}$와 대화 *dialogue*를 함께 LLM 모델 *Llama*2에 입력한다. 출력된 초기 응답 $response_{init}$을 질의로 하고 (0)단계에서 검색된 *top*-k개 문서 $doc_{top\text{-}k}$를 대상으로 검색기 *retrieval*를 이용해 재순위화를 수행한다. 재순위화된 결과에서 *top*-1 문서 doc_{top1}를 *threshold*로 설정하고, *threshold* 이하의 문서를 추출한다. 마지막으로 (3)단계와 같이 *dialogue*와 *threshold* 이하의 문서 $doc_{top\text{-}k} <= threshold$를 피드백 문서로 사용해 *Llama*2에 입력하여 최종 응답 *response*를 출력한다.

3.1 검색 단계

문서 그라운딩된 대화 시스템 태스크 수행을 위해 주어진 대화 *history*와 상위 연관된 문서 $doc_{top\text{-}k}$ 검색을 그림 1의 (0)단계와 같이 선행하였다. 또한 초기 응답 $response_{init}$을 이용해 피드백 문서를 선별하기 위해 검색을 수행하였다. 본 연구에서는 검색을 위해 DPR을 사용하였다. DPR은 질의와 검색 대상 문서를 가각 인코딩하고, 인코딩된 결과를 내적하여 유사도를 계산하는 방법으로 질의에 대한 연관된 문서를 검색한다. 인코딩되기전에 모델 입력 형태인 $input_ids \in \mathbb{V}^{max_len}$로 변환하고, 변환된 $input_ids$가 식 (1)와 같이 각각 질의 인코더 $E_Q(\cdot)$, 문서 인코더 $E_P(\cdot)$를 통과 내적하여 유사도 $sim(q,p)$가 계산된다. 이 때, $\mathbb{V}$는 모델의 vocab 사이즈를 의미하고, max_len은 모델 최대 입력 길이를 의미한다.

$$sim(q,p) = E_Q(q)^T E_P(p) \tag{1}$$

(0)단계에서의 검색은 q가 대화 *dialogue*를 의미하고, p가 문서 집합 $D = doc_1, doc_2, ...doc_N$를 의미한다. (2)단계에서의 검색은 q는 초기 응답 $response_{init}$를 의미하고, p가 대화 *dialogue*에 대한 검색 결과 $doc_{top\text{-}k}$를 의미한다.

$$output_{last} = Llama2(prompt_{last}) \tag{2}$$

3.2 초기 응답 생성 단계

본 연구에서는 피드백 문서를 선별하기 위해 Llama2 모델이 생성한 초기 응답 $response_{init}$을 이용하였다. 초기 응답 $response_{init}$ 생성을 위해 (0)단계에서 검색된 *top*-1 문서 $doc_{top\text{-}1}$와 대화 *diaglogue*를 *Llama*2 모델에 입력하였다. 모델 입력 형태는 표1의 초기 응답 생성 (1) step의 prompt를 사용해 변환하였고, 이를 이용해 In-Context learning을 수행하였다. 식(3)와 같이 (0)단계에서 검색된 *top*-1 문서 $doc_{top\text{-}1}$와 대화 *diaglogue*는 입력 프롬프트 $prompt_{init}$로 변환되어 *Llama*2 모델에 입력되고 출력된 $output_{init}$은 후처리되어 자연어 형태인 $response_{init}$으로 변환된다.

$$output_{init} = Llama2(prompt_{init}) \tag{3}$$

3.3 피드백 문서 선별 단계

피드백 문서 선별은 그림1의 (2)단계와 같이 수행된다. 사용자 마지막 질의에 대해 적절한 응답을 생성하기 위해 (1)단계에서 생성된 초기 응답 $response_{init}$과 (0)단계에서 검색된 문서 $doc_{top\text{-}k}$를 이용하였다. 먼저, 초기 응답 $response_{init}$과 검색된 문서 $doc_{top\text{-}k}$을 각각 q, p로 사용해 검색 단계와 동일한 DPR

Retrieval Model	MRR (%)	R@1 (%)	R@5	R@10	R@30	R@50	R@100
Dens Passage Retrieval	65.63	53.55	81.39	87.14	93.80	95.16	97.28
Re-Ranking	71.78	62.48	83.36	89.41	93.80	-	-

표 2. 검색 시스템 실험 결과

Dataset type	Model	F1 Score	Rouge-L Score	SacreBleu Score	Meteor Score	Total Score
Gold Set	Baseline (k=5)	18.82	15.26	3.17	18.68	55.93
	Proposed	18.81	15.15	3.21	18.42	55.59
Retrieved Set	Baseline (k=5)	15.58	12.23	2.49	15.05	45.35
	Proposed	**15.78**	**12.53**	**2.69**	**15.23**	**46.23**

표 3. 문서 그라운딩된 대화 시스템 Zero-Shot 태스크 실험 결과

을 수행해 초기 응답 $response_{init}$의 검색된 문서 $doc_{top\text{-}k}$에 대한 유사도를 계산한다. 유사도를 기반으로 $doc_{top\text{-}k}$을 재순위화하고, (0)단계에서의 *top*-1 문서 $doc_{top\text{-}1}$를 *threshold*로 설정하였다. 초기 응답과 유사도를 계산하였을 때, 특정 *threshold* 이하의 유사도를 보인 문서는 응답 생성에 반영이 적게 되었다는 의미이기 때문에 최종 응답 *response* 생성 시에 추가적으로 입력할 수 있도록 *threshold* 이하의 문서들을 피드백 문서로 선별하였다.

3.4 최종 응답 생성 단계

최종 응답 생성은 그림 1의 (3)단계와 같이 수행된다. 질의에 대한 적절한 응답을 위해 그림1의 (2)단계에서 선별된 피드백 문서들과 대화 *dialogue*가 표1의 최종 응답 생성 (3) step의 프롬프트 $prompt_{last}$로 변환되어 입력된다. *Llama*2 모델에 입력된 프롬프트 $prompt_{last}$는 식 (2)와 같이 *In-Contextlearning*을 수행해 $output_{last}$가 출력되고, 후처리되어 자연어 형태인 *response*로 변환된다.

4. 실험

4.1 태스크 정의

본 연구에서 수행한 태스크는 문서 그라운딩된 대화 시스템으로 대화 기록에 근거해 연관된 문서를 선행 검색하고, 사용자 마지막 질의에 대해 검색된 문서가 그라운딩된 응답을 생성하는 것을 목적으로 한다. 대화에 문서가 매핑되어 있는 MultiDoc2Dial[18] 데이터셋을 사용하였고, Seen-data의 test 셋을 이용해 응답 생성 성능을 측정하였다. 응답 생성 전 선행되는 검색 과정은 Baeline과 비교 없이 DPR 모델과 DPR 모델에서 출력된 결과에 대해 Recall, MRR-Score를 이용해 Re-Ranking 성능을 측정하였다. 응답 생성 과정은 Llama7B 모델을 기반으로 실험하였고, In-Context learning 실험을 위해 Zero-Shot을 수행하였다. 본 연구에서 제안한 모델과 비교를 위해 검색된 *top*-1 문서를 같이 입력하는 $response_{init}$ 모델을 Baseline으로 설계하였다, token-level F1 Score와 Rouge-L, SacreBleu matrics, Meteor-Score를 사용하였고, 네 지표의 합산 점수를 기준으로 모델별 성능을 비교하였다.

4.2 검색 시스템 실험 결과

검색 단계는 DPR과 Re-Ranking을 이용하였고, 표 2와 같은 성능을 보였다. Re-Ranking은 DPR 결과의 top-30의 문서를 기준으로 수행하였기 때문에 top-30 까지의 결과를 기재하였다.

4.3 응답 생성 실험 결과

응답 생성 실험 결과는 표 3와 같다. Data type의 Gold Set은 gold data를 포함한 실험의 결과이고, Retrieved Set은 검색된 결과에 기반한 실험의 결과이다. Zero-Shot 기반으로 실험을 하였을 때, 응답 생성 성능이 저하된 것을 확인할 수 있다. Baseline의 경우, 검색된 top-5 문서를 참고하였고, Proposed 모델은 검색된 top-5 문서 중 피드백 선별 문서 선별을 통해 instance마다 다른 개수의 문서가 참고되었다. Gold Set 실험의 경우, 약 0.4점의 Total Score의 차이로 Baseline의 성능이 우수하지만 Retrieved Set 실험은 약 0.9점의 Total Score의 차이로 제안 모델의 성능이 우수한 것을 확인할 수 있었다. Retrieved Set의 성능에서 향상된 것을 통해 초기 응답과 검색된 문서를 이용해 피드백 문서를 선별한다면 검색된 문서가 정답이 아니더라도 LLM이 주어진 문서들에 기반해 적절한 응답을 생성할 수 있다는 것을 확인하였다.

5. 결론 및 향후 연구

Llama2 언어 모델에 기반해 문서 그라운딩된 대화 시스템 태스크를 수행하였다. 본 연구에서 제안한 피드백 선별 기법 적용

이 성능 향상에 영향을 줄 수 있는지 확인하고자 fine-tuning을 수행하지 않고, Zero-Shot 기반 In-Context learning을 수행하였다. 그 결과, 기존 연구들의 fine-tuning 결과와 비교해 응답 생성 성능이 저조한 것을 확인할 수 있지만 같은 Zero-Shot 기반 In-Context learning 환경에서는 검색된 top-1 문서를 입력한 Baseline과 비교해 제안 모델이 검색된 문서 환경에서 성능이 향상된 것을 확인할 수 있었다. 실제 환경과 동일한 검색된 문서에 기반해 응답을 생성하였을 때, 초기 응답과 검색된 문서를 이용해 피드백 문서를 선별해 사용한다면 응답 생성 성능이 향상될 수 있음을 확인하였다.

태스크에 대해 fine-tuning한 기존 연구들과 비교해 Zero-Shot 기반 In-Context learning 환경의 응답 생성 성능이 저조한 것을 확인하였다. 향후, prompt 개선을 통해 성능 향상을 도모하고, 초기 응답 생성을 수정해 더 적합한 응답을 생성할 수 있는 방법을 연구하고자 한다.

6. 감사의 글

이 논문은 2023년 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2021-0-02068, 인공지능 혁신 허브 연구 개발)

This work was supported by Institute of Information & communicaions Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-02068, Artificial Intelligence Innovatio Hub)

참고문헌

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., pp. 5998–6008, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[2] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds., pp. 7871–7880, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.703

[3] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, Vol. 21, pp. 140:1–140:67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html

[4] V. Karpukhin, B. Oguz, S. Min, P. S. H. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih, "Dense passage retrieval for open-domain question answering," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., pp. 6769–6781, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.emnlp-main.550

[5] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," 2023.

[6] J. Scheurer, J. A. Campos, T. Korbak, J. S. Chan, A. Chen, K. Cho, and E. Perez, "Training language models with language feedback at scale," *CoRR*, Vol. abs/2303.16755, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.16755

[7] Y. Wang, W. Zhong, L. Li, F. Mi, X. Zeng, W. Huang, L. Shang, X. Jiang, and Q. Liu, "Aligning large language models with human: A survey," *CoRR*, Vol. abs/2307.12966, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2307.12966

[8] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., pp. 4171–4186, 2019. [Online]. Available: https://doi.org/10.18653/v1/n19-1423

[9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, Vol. abs/1907.11692, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[10] K. Clark, M. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: pre-training text encoders as discriminators rather than generators," *CoRR*, Vol. abs/2003.10555, 2020. [Online]. Available: https://arxiv.org/abs/2003.10555

[11] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *CoRR*, Vol. abs/2005.14165, 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

[12] OpenAI, "Gpt-4 technical report," 2023.

[13] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023.

[14] S. Lin, J. Hilton, and O. Evans, "Truthfulqa: Measuring how models mimic human falsehoods," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds., pp. 3214–3252, 2022. [Online]. Available: https://doi.org/10.18653/v1/2022.acl-long.229

[15] N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano, "Learning to summarize with human feedback," *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/1f89885d556929e98d3ef9b86448f951-Abstract.html

[16] R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders, X. Jiang, K. Cobbe, T. Eloundou, G. Krueger, K. Button, M. Knight, B. Chess, and J. Schulman, "Webgpt: Browser-assisted question-answering with human feedback," 2022.

[17] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," *NeurIPS*, 2022. [Online]. Available: http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html

[18] S. Feng, S. S. Patel, H. Wan, and S. Joshi, "Multidoc2dial: Modeling dialogues grounded in multiple documents," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds., pp. 6162–6176, 2021. [Online]. Available: https://doi.org/10.18653/v1/2021.emnlp-main.498