인공지능 언어모델의 상식적 판단

송상헌(고려대학교)

1. 들어가며

우리는 일상생활에서 ‘상식’ 또는 ‘상식적’이라는 말을 단어를 의외로 자주 사용한다. 표준국어대사전에 따르면 상식(常識)은 아래와 같이 정의되며, ‘보통지식’이라는 말과 유의어 관계를 가진다.

(1) 사람들이 보통 알고 있거나 알아야 하는 지식. 일반적 견문과 함께 이해력, 판단력, 사리 분별 따위가 포함된다.

위와 같은 정의는 상식에 대한 가장 기본적인 두 가지 사실을 담고 있다. 첫째는 ‘보통성’으로 적어도 성인이라면 누구나 알고 “있어야” 마땅할 정도로 우리가 사는 세계에 대한 보편적 지식을 말한다. 이는 과학, 법학, 역사학 등의 ‘전문지식’과 구분되는 것이다. 둘째는 단순한 지식이 아니라 ‘사고력’의 영역까지 확장된다는 것이다. 즉, 무엇을 알고만 있는 것이 아니라 주어진 상황에서 어떠한 일반적 결론을 도출할 수 있어야 ‘상식적 판단’을 갖추었다고 평할 수 있다. 이러한 맥락에서 우리말샘에서 발췌한 아래 예문들을 살펴 보기로 하자. 아래의 예문들은 실제로 우리가 상식에 대해서 암묵적으로 상정하고 있는 개념들을 잘 나타내고 있다.

(2) ㄱ. 그 뒤에는 이일보다도 한 수가 위라는 신립이 또 있으니 두 장수가 문경 새재쯤 지킬 것은 당연한 상식 이하의 상식이었다.≪박종화, 임진왜란≫
ㄴ. 내 질문에 현구는 너무나 상식적인 물음이란 듯 멀뚱한 얼굴로 나를 올려다보았다.≪김원일, 노을≫

(3) ㄱ. 그 양반이 태남이에게 보여 주던 그 깊은 관심과 자애는 상식으로는 설명할 수 없는 거여서 그 당시엔 노망으로 돌렸고….≪박완서, 미망≫
ㄴ. 세 살 먹은 어린애가 본다 해도 이 사건은 상식적으로 말도 안 돼요.
≪이호철, 문≫

(4) ㄱ. 그것은 병원에서 진단할 것이지만 워낙 사건이 상식 밖의 일이라서 저의 질문에 대한 답변을 신용하기 위해서입니다.≪김용성, 리빠똥 장군≫
ㄴ. 거기다가 아무리 확신범의 특질을 감안한다 할지라도 범행 동기는 너무도 상식과 먼 거리에 있었다.≪이문열, 사람의 아들≫

(5) ㄱ. 그는 결점을 늘어놓아 상대편 기분을 해치는 것만큼 몰상식은 없다고 생각했다.
ㄴ. 그는 그녀가 공연장에서 떠들고 웃는 몰상식한 행동을 보이자 그만 나와 버렸다.

첫째, 상식은 교육을 통해서 이루어지는 것이 아니다. 특정 사회의 구성원들이 별도의 교육이 없이도 암묵적으로 공유하고 있는 지식 체계이다. 이는 상식은 책에 기술되어 있거나 문서로 정리되어 배포되는 것이 아니라는 의미이다. 그러함에도 누구나 그렇게 알고 있고 그렇게 생각할 것으로 기대되는 바이다. 상식이 가지는 이러한 특성이 잘 드러나는 예문이 (2)에 제시되어 있다. (2ㄱ)과 같은 전시 상황에서 장수(將帥)로서의 상식이 있다면 '문경 새재'를 요충지로 생각하는 것이 당연하고, 따라서 별도의 명령으로 하달해야 할 필요성이 없다. 누군가 아주 당연한 내용에 대해서 질문을 하게 되면 질문을 받는 사람은 (2ㄴ)에 묘사된 것처럼 당황하게 된다.

둘째, 위와 같은 근거에서 상식의 가장 일차적인 기능은 설명을 위한 도구이다. 우리가 사는 세계에서 벌어지는 여러 사건과 현상을 이해하고 설명하기 위해서 우리는 상식이라는 공통된 잣대를 사용하는 것이다. (3)의 예문에서 나타나는 바와 같이 상식적 판단 능력은 그 사람의 지적 수준이 정상적인지를 가늠하는 척도가 될 수도 있다. 누군가가 상식에서 벗어나는 언행을 한다면 그 사람은 '노망'이 들었거나 '세 살 어린이'만도 못한 지적 능력을 가졌다는 뜻이다. 이는 상식적 판단 능력을 보면 '지능'이 제대로 작동하고 있는지를 알 수 있다는 의미이기도 하다.

셋째, 위와 같은 이유에서 상식에는 그 사회 구성원들의 어떠한 기대치가 반영되어 있다. (4ㄱ)에서 예시된 바와 같이 어떠한 사건이 상식 안의 일이냐 그렇지 않은가와 같은 이분지 판단을 내릴 수 있다. 또는 (4ㄴ)에서 묘사된 바와 같이 거리와 같은 '정도성'의 문제로 상식을 활용하기도 한다. 즉, 준거 판단을 위한 일종의 잣대로 기능을 할 수 있다는 것이 상식이 가진 중요한 역할이다. 그래서 그 기준에 부합하지 않는 사건과 상황을 특이점을 가진 것으로 주목한다. 이는 우리가 세계의 여러 현상 등을 다루는데 아주 효과적인 방편이다. 상식적인 것의 범주를 어느 정도 정해놓고 그 기준치에서 벗어나는 것에 대해서만 문제 해결의 에너지를 집중하면 되기 때문이다.

넷째, 상식은 때로 '윤리', '도덕', '예의범절'과 유사한 부류로 묶인다. 대표적으로 상식에서 파생된 '몰상식'이라는 단어의 용례에서 그러한 개념을 살필 수 있다. (5ㄱ)에서 보이듯이 타인의 결점을 공공연히 밝히는 행동은 사회통념에 어긋나는 것으로 이 역시 명시적으로 학습되지 않았어도 문화인 또는 교양인이라면 상식적 판단이다. 공공의 장소에서 소란을 피우는 (5ㄴ)의 상황도 상식적 판단의 영역에 드는 일이다. 우리가 흔히 '가정교육'이라고 하는 말에서 알 수 있듯이 정규 교육기관에서 가르치는 내용이 아니지만 사회성을 갖춘 사람이라면 응당 취해야 하는 행동 규범 역시 상식의 테두리 안에 있다. 실제로 우리말샘의 규범 정보에 따르면 '몰상식하다'와 '못되다'를 함께 쓸 수 있다고 되어 있다(행정 용어 순화 편람, 1993년 2월 12일). 즉, 많은 경우 '상식'과 '윤리'는 동전의 양면과도 같다.

상식이 가진 특성을 실제 말뭉치의 용례를 통해 조금 더 살펴보자. 아래 (6-8)은 연세 20세기 한국어 말뭉치에서 발췌한 예문들이다. 여기에서 20세기 말뭉치를 대상으로 하는

것도 이유가 있는데, 상식은 단기간에 형성되는 것이 아니라 오랜 기간에 걸쳐서 역사적으로 형성되는 것이고 또한 한번 형성된 상식은 다시 그 사회에서 오랫동안 영향력을 발휘한다. 아래에서 보듯이 60년대에서 80년대에 통용되던 상식이 몇 세대가 지난 지금까지 유효하게 작용을 한다.

(6) ㄱ. 규율 부장은 다른 학생들의 모범이 되어야만 한다는 상식 정도는 나도 알고 있었다. ≪세계는 넓고 할 일은 많다[내 사랑하는 젊은이들에게]≫ 1980년대
ㄴ. 무슨 말씀을요.....보스께서 먼저 타시는 게 상식 아니겠습니까? ≪사계의 후조(상)≫ 1970년대
ㄷ. 더구나 협박에서 풀려나는 댓가로 요구한 금액이 겨우 십만원이란 것도 상식 밖의 일이었다. ≪아홉켤래의 구두로 남은 사내≫ 1970년대

(7) ㄱ. 어두운 화제보다는 밝은 화제를, 내 얘기보다 상대방 얘기나 쌍방간에 공명할 수 있는 화제를 택해 가지고 알기 쉬운 표현으로 말하는 것이 상식있는 사람의 취할 태도입니다. ≪에티켓 선생(상)≫ 1960년대
ㄴ. 교양이란 거죽에만 살짝 덮여졌을 뿐이고, 그 속에는 온통 상식 이하의 것들이 도사리고 있다. ≪인간적인 진실로 인간적인≫ 1970년대
ㄷ. 소위 해방군이라고 자처라는 저들이 수준이 상식 이하라는 사실에 놀라지 않을 수 없었읍니다. ≪끝이 없는 이 길을≫ 1970년대

(8) ㄱ. 그 상품을 입수(入手)하자면 그만한 대가를 치뤄야 하는 것이 당연한 일이요, 영화를 구경하자면 관람료를 지불해야 하는 것은 상식 이하의 상식이 아니고 무엇이던가. ≪비석과 금강산의 대화≫ 1960년대
ㄴ. 안 해도 될 일을 해야 할 때는 무슨 속셈이 있다고 보는 게 상식 아니겠소? ≪하오의 戀歌≫ 1960년대

(6)에서 제시된 용례는 앞서 설명한 상식의 일반적인 특성을 다시 밝히고 있다. 상식은 명문화 혹은 성문화되어 있지 않다(6ㄱ). 상식은 별도의 교육이나 학습이 없이도 심지어 조직폭력배라도 공유하고 있는 지식 체계이며(6ㄴ), 그 범위를 정도성으로 나타낼 수도 있다(6ㄷ). (7)의 예시에서 보이는 바와 같이 상식은 종종 '에티켓', '교양', '윤리' 등과 같은 범주에서 논의된다. 상식이 가지는 한 가지 또 다른 특징이 (8)에서 예시되어 있는데, 흔히 세계의 지식을 패턴화한다고 한다. 이는 유추와 비슷한데 일련의 관찰을 통해 공통의 법칙을 발견하여 이를 하나의 명제로 도출하는 과정이다. 이 흐름은 이른바 가추(假推) 추론의 범주 안에 포함된다.

같은 방식으로 세종 말뭉치에서 출현한 용례를 이해해 보자. (9)의 예문들은 상식이 '전문지식'에 대별되는 일반인은 '보통지식'이라는 점을 나타낸다. (10)의 예문들은 상식과 윤리가 상호 밀접한 관련성을 가진다는 증례에 해당한다.

(9) ㄱ. 그러므로, 우리의 역사는 분석의 역사가 아니라 종합의 역사요, 전문적 역사가 아니라 상식적 역사다. ≪역사와 민족≫
ㄴ. 정보통신 인프라라면 사무용 건물에나 적용될 것이라는 것이 일반인의 상식이다. ≪월간중앙 5월호≫
ㄷ. 저자는 매우 복잡한 현상이나 전문용어로 기술될 실험장치들까지도 상식적인

수준에서 이해시키고자 한다. ≪천리안의 자연과학 서적 서평≫

(10) ㄱ. 그는 삶이나 건강이나 모두 평범한 상식과 원칙 속에 바른 길이 있다는 것을 몸소 실천하고 보여 주며 살아가고 있다. ≪좋은생각 1999년 12월호≫

ㄴ. 김밥 찌꺼기, 깡통 따위는 주위의 눈치도 보지 않고 한강에다 던져 넣는 몰상식한 시민들이 적지 않다. ≪조선일보 과학(93)≫

ㄷ. 난 도대체 어떻게 그런 몰상식한 행동, 비양심적인 태도로 나올 수 있는가 도저히 이해할 수가 없습니다. ≪잠자는 갈매기≫

ㄹ. 상식적으로 용납할 수 없는 일을 범하거나 통상적인 윤리로 받아들이기 어려운 일을 묵과하는 태도는 명철한 생활인의 자세일 수가 없다. ≪인간과 사회 -전통윤리와 현대풍조의 갈림길에서≫

이상과 같은 상식에 대한 전형적인 특성 외에도 (11) 및 (12)에서 부가적인 특성이 관찰된다. (11)의 예문은 앞서 (8)에서 보인 바와 같이 어떠한 관찰 사실을 바탕으로 유추에 의한 개념 확장으로 상식이 기능을 할 수 있다는 점을 드러낸다. (12)의 예문은 상식의 또 다른 특성을 보인다. (12ㄱ-ㄴ)에서 설명된 바와 같이, 상식은 고정불변의 것이 아니며 그러하기에 오류 가능성을 내포한다는 것을 나타낸다. '여성을 열등하다'고 보는 것은 오랫동안 지속되어 온 잘못된 상식이기에 폐기되어야 할 대상이 된다(12ㄷ). 우리가 알고 있는 의약 상식 가운데도 실제 과학적 근거가 희박한 것들이 상당하다(12ㄹ).

(11) ㄱ. 평소 안경을 끼던 사람은 벗고 안 끼던 사람은 끼는 것이 도피의 상식이다. ≪광야의 끝에서≫

ㄴ. 남성이 만들어 낸 '남성=사회역할, 여성=가정역할'이라는 효율적인 역할 구분은 마땅히 여자는 여자답고 남자는 남자다워야 한다는 상식 세계의 터전을 다져 놓았다. ≪여성의 일곱가지 콤플렉스≫

(12) ㄱ. 어제의 상식만으로는 오늘의 문제를 풀 수 없다. ≪정보교육≫

ㄴ. 우리들 사이에 통용되어온 상식에 어떤 모순이 있지 않은가? ≪사회를 보는 논리≫

ㄷ. 그리하여 어릴 때부터 가족, 학교, 매스컴 등을 통해 여성이 열등하다는 것은 상식으로 굳어 갔다. ≪여성의 일곱가지 콤플렉스≫

ㄹ. 잘못된 의약 상식과 보신에 대한 집착이 결합된 결과라고 할 수 있다. ≪사회를 보는 논리≫

지금까지 용례를 통해 살펴본 상식의 일반적인 특성을 정리하면 다음과 같다. 상식은 책이나 문서를 통해 정리된 명문화/성문화된 전문적인 지식이 아니다. 오랜 역사적 배경을 통해 문화인이라면 누구나 암묵적으로 공유하고 있는 내용이다. 따라서, 그 내용을 궁금해하거나 혹은 그에 반하는 행동을 하는 것은 분별력 있는 사람에게는 이상하게 받아들여지게 된다. 이처럼 상식적 판단은 예상 가능한 범주 안에 들어야 하므로, 우리 사회의 여러 현상에 대한 설명의 도구로서 기능한다. 또한 많은 경우 상식은 윤리도덕, 예의범절, 규범질서 등과 같은 선상에서 취급된다. 때에 따라 관찰을 통해 유추적 패턴을 만들어 상식적 판단을 생성할 수도 있다. 이 과정이 항상 완벽한 것은 아니기 때문에 상식은 얼마든 틀릴 수 있고 또 그러기에 수정될 수 있다.

이상과 같이 상식에 대해 세세한 설명을 거듭한 까닭은, 최근 인공지능 분야에서 상식이 가진 이와 같은 특성이 매우 중요한 과제로 부상하고 있기 때문이다. 이어지는 절에서 인공지능의 상식 추론에 대한 동향을 개괄하고 그 중요성과 발전 방향에 대해 개괄하겠다.

2. 언어 인공지능과 상식

2.1 연구의 흐름

최근 언어 인공지능의 개발과 활용에 있어서는 GPT3와 같은 초거대 학습 데이터를 활용하는 생성 모델이 각광을 받고 있다. 이러한 분위기에 편승하여 인공지능 언어모델에 대한 몇몇 소문이 돌고 있다. 대표적으로 GPT3가 사람보다 뛰어날 정도로 완벽성을 갖추고 있다거나 GPT3의 서사구조가 연구 대상이 된다거나 하는 주장이다. 물론 최근의 생성 모델이 비약적인 발전을 거듭한 까닭에 문장 생성을 매끄럽게 잘하는 것은 주지의 사실이다. 그러나 '인간을 뛰어넘는다'라는 정량지표를 객관적으로 어떻게 확립하였는지 모르겠다. 또한 문장 단위를 넘어 서사의 구조를 생성하는 것은 아직 쉬운 과제가 아닌데, 저러한 주장을 무슨 근거로 그렇게 쉽게 하는지도 모르겠다.

다만 본고에서 언급하고자 하는 바는 그 엔진을 이해할 필요가 있다는 것이다. 비유적으로 전기자동차를 생각해보자. 전기자동차 엔진은 내연기관이 아니라 전동기로 힘을 발생시키고 그래서 휘발유 대신에 전기를 사용한다. 전기자동차가 그 이전까지 존재하였던 자동차와 가장 크게 차이가 나는 점은 바로 이 엔진의 구조와 동력원이다. 우리가 전기자동차에 관해서 설명하고자 한다면 이러한 특성에서부터 설명을 이끌어가야 순서와 이치에 맞을 것이다.

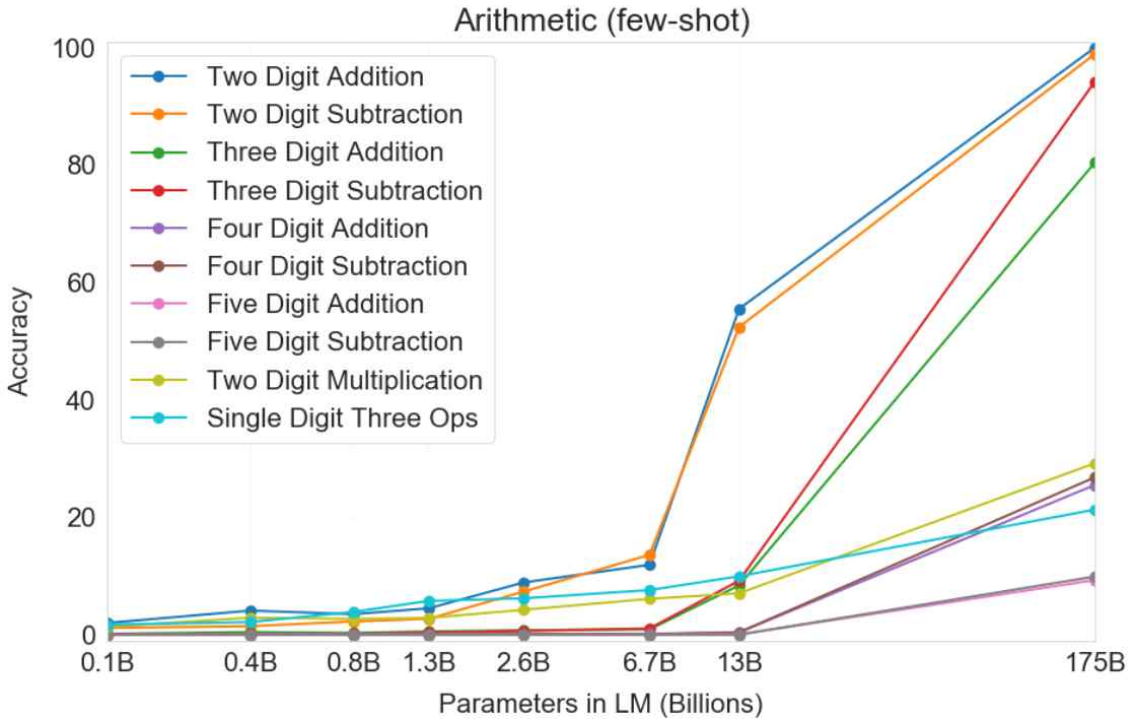
최근의 몇 년간의 언어 인공지능의 발전 속도는 자동차가 지난 수십 년간 발전해 온 속도를 상회한다. 그렇다면 전기자동차의 엔진과 마찬가지로 최근의 기술 도약에 해당하는 그런 괄목할 변화는 무엇일까? 이 질문에 대해서는 여러 답변이 있을 수 있겠으나 본고는 'Beyond the language! Beyond the surface language!'라고 답을 하고자 한다.

이전까지의 전산언어자원은 말 그대로 언어 자체를 처리하기 위해 쓰였다. 형태분석은 다른 언어 처리 모형 예컨대 정보 검색이나 구문 분석을 하기 위한 전처리 단계로 기능을 하였다. 구문분석을 하는 가장 큰 이유는 문장에 대한 의미표상을 구하기 위해서였으며, 그렇게 구해진 의미표상은 기계번역의 입력 자료로 활용이 되었다. 즉, 일련의 전산언어자원이 물고 물리는 관계였지만, 결국 자연어 자체를 이해하고 처리한다는 점에서는 큰 틀에서 같았다. 그런데 최근의 언어모델은 언어 이상의 것, 다시 말해 언어가 담고 있는 그 이상의 정보와 지식을 이해하고 생성하는 방향으로 가고 있다.

이러한 현황을 잘 다루고 있는 최근 저작으로 네 편의 논문을 소개한다. 이 저작들은 딥러닝 언어연구의 최신 동향에 대해 문의를 하시는 분들에게 필자가 추천 드리는 논문들이기도 하다.

첫 번째 논문은 GPT3가 소개된 Brown et al.(2020)이다. 이 논문은 매우 많은 저자를

포함하며 그만큼 많은 내용과 다층적인 실험을 포괄하고 있다. 그 가운데서 언어 인공지능의 발달 방향과 관련하여 가장 눈여겨볼 항목은 아래 <그림 1>이다. 이 그래프는 언어 인공지능의 일종인 GPT3에게 사칙연산을 시켜본 그 결과를 나타낸다. 즉, 언어모델을 매우 방대한 크기로 만들었더니, 그 결과 두 자릿수의 덧셈뺄셈까지는 완벽하게 수행함을 발견하였다는 것이다. 아주 쉽게 말해, 국어공부를 엄청나게 시켰더니 산수까지 할 수 있게 되었다는 뜻이다. 이 결과가 처음 공표가 되었을 때, 학자들 사이에서도 논쟁이 있었다. 두 자릿수 덧셈과 뺄셈이 그렇게 가능한 것은 학습데이터가 매우 크기 때문에 그 어딘가에 그러한 덧셈뺄셈의 예시가 포함되어 있기 때문이지 실제 사칙연산의 법칙까지 학습한 것으로 볼 수는 없다는 견해가 있었다. 필자도 처음 이 그래프를 접했을 때는 이와 유사한 의견이었다.



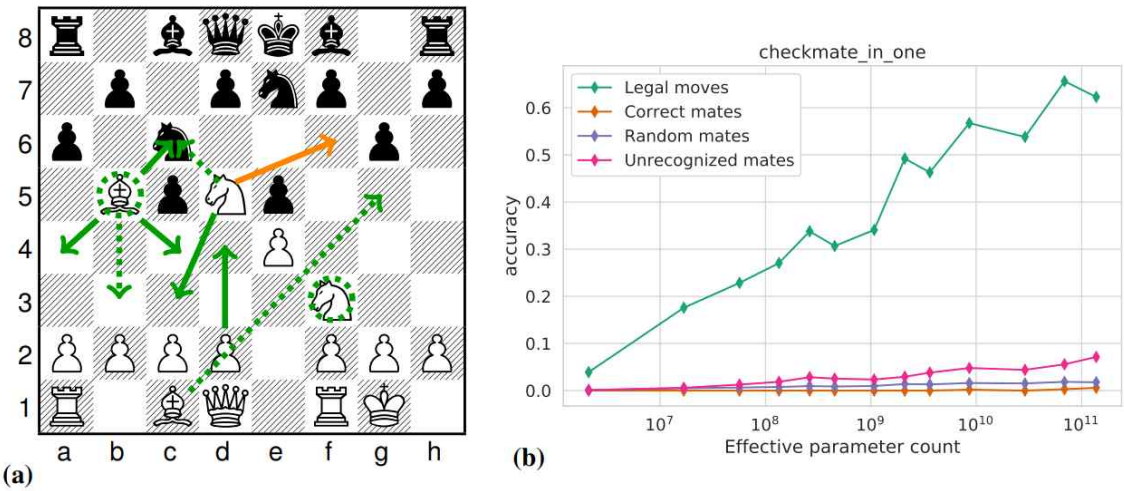
<그림 1> GPT3의 사칙연산 (Brown et al. 2020, 22쪽)

다음의 두 편의 논문 Wei et al. (2021)과 Kojima et al. (2022)은 최근의 이른바 zero-shot 학습에 대한 것이다. 최근의 대규모 언어모델들은 이른바 fine-tuning 없이도 다양한 과업을 실제 수행할 수 있게 발전을 거듭하였다. 이러한 대규모 모델들은 인공지능에게 예시나 지침을 제공하는 '프롬프팅(prompting)'을 어떻게 쓰느냐에 따라 진일보한 과업 성능을 보여 준다. 예컨대, 하단의 <그림 2>는 "Let's think step by step"이라는 문구가 프롬프트로 제시되었을 때 모델의 실제 성능이 크게 개선될 수 있다는 것을 나타낸다. 이는 현재의 대규모 언어모델이 단순한 'learner' 학습자가 아니라, 논리 이해 능력을 갖춘 'reasoner' 즉, 추론자가 될 수 있다는 것을 함의한다. 프롬프트가 무엇이냐에 따라 문제 해결에 대한 접근법을 자체적으로 미세 조정할 수 있다는 의미이기 때문이다.

No.	Template	Accuracy
1	Let's think step by step.	**78.7**
2	First, (*1)	77.3
3	Let's think about this logically.	74.5
4	Let's solve this problem by splitting it into steps. (*2)	72.2
5	Let's be realistic and think step by step.	70.8
6	Let's think like a detective step by step.	70.3
7	Let's think	57.5
8	Before we dive into the answer,	55.7
9	The answer is after the proof.	45.7
-	(Zero-shot)	17.7

<그림 2> 프롬프트에 따른 정확도 향상 (Kojima et al. 2022, 8쪽)

끝으로 흔히 BIG Bench라고 불리는 평가 체계를 제안한 Srivastava et al. (2022)를 참조할 수 있다. 대표적으로 아래 <그림 3>에서 제시된 체스의 한 장면을 통해 언어 인공지능의 발전 방향을 이해할 수 있다. 언어 인공지능을 평가할 때 단순히 언어 데이터만 대상으로 하는 것이 아니라, 체스에서 체크메이트 상황을 판별할 수 있는지를 평가의 척도로 활용하는 것이다. 우리가 흔히 '지능'이 뛰어난 사람을 떠올릴 때 많이 연상할 수 있는 예시가 '체스' 능력일 것이다. 즉, 특정한 언어 인공지능이 세계의 지식을 제대로 학습하여 추론 능력을 제대로 갖추었다면 체스에 관한 판단 능력도 적어도 일반인을 상회하는 만큼의 수준이 되어야 우리의 기대치에 맞을 것이다.



<그림 3> 체크메이트 판별 (Srivastava et al. 2022, 21쪽)

이러한 흐름은 사실 우리가 일반적으로 알고 있는 바에서 크게 벗어나는 것이 아니다. 언어 능력이 좋으면 학습 능력이 덩달아서 좋아지고 그에 따라 추론 능력까지 상승하는 것은 우리가 생각하는 범위에서 일견 타당한 연상 작용이다. 모국어에 대해서 충분한 어휘 지식을 획득하고 나면 그 어휘가 함의하는 개념에 대한 이해도가 높아져서 사회 혹은

과학 과목에 대한 학습을 더 쉽게 할 수 있는 것은 충분히 가능한 시나리오이다. 또한 독서 능력을 보다 집중적으로 훈련받은 학생 혹은 그만큼 습관이 갖추어져 있는 학생이 그만큼 전 과목의 학습 능력이 좋다는 것 역시 여러 연구를 통해 증명된 사실이다. 그래서 아래 <그림 4>와 같은 저작들이 있다(정도상 2017; 신성일 2014).

<그림 4> 모국어 능력을 성적 향상의 선결 조건으로 강조한 저서들

언어에 대한 이해 능력은 결국 추론 능력까지 확장될 수 있다는 것이 위와 같은 저작들이 가지는 공통된 가정이다. 언어 인공지능과 상식을 연계하려는 시도 역시 이러한 맥락이다. 서론에서 예시를 통해 본 바와 같이 상식 자체는 물론 언어가 아니다. 그러나 언어로 포장되어 있다. 한때, 자연어 연구의 중요한 단위인 정보 구조(information structure)를 정보 포장(information packaging)이라는 용어로 불린 바 있는데(Engdahl and Vallduví, 1996), 이는 언어가 어떠한 정보를 담아서 전달하는 매개체의 기능을 부각한 것이다. 언어라는 형식에 포장되어 전달되는 정보가 세대를 이어 통시적 차원이 되었을 때, 상식은 그 대표적인 대상이 될 것이다.

2.2 인공지능 평가로서의 상식적 판단

서론에서 여러 예시와 함께 설명한 바와 같이, 상식의 심리학적 정의는 특정 사회의 구성원들이 특별한 교육을 받지 않아도 암묵적으로 공유하는 지식 체계이다. 우리에게는 어쩌면 이와 같은 정의보다 취업 등에서 사용하는 이른바 “일반상식”이 더 친숙한 개념일 수 있다. 그러나 적어도 최근의 인공지능 연구에서 주안점을 두고 있는 상식은 이러한 일반상식과는 명확한 차이가 있다. 예컨대, 속초의 특산물이 무엇인가 하는 것은 속초에 관한 별도의 공부를 해야 알 수 있는 지식이다. 비슷한 예로 TV 예능프로그램에서는 상식 퀴즈를 하면서 주로 수도(capital) 맞추기를 한다. 물론, 미국이나 일본과 같은 주요 국가

의 수도까지는 기본적인 교육을 받은 사람의 상식이라고 할 수 있을지 모르겠다. 그러나 코트디부아르와 같이 대부분에게 나라 이름조차 생소한 경우 세계 지리 등에 관한 관심으로 따로 공부를 한 사람이 아닌 한, 수도 이름을 알기가 어렵다.

인공지능 연구에서 활용되는 상식의 특성을 잘 대변할 수 있는 사례로 도구를 들 수 있다. 이는 심리언어학 등의 실험에서 사용하는 도구 논항 등과 유사한 개념이다(Rissman et al. 2015). 예컨대 우리가 손톱을 깎을 때 물론 가위나 칼로 깎을 수는 있다. 그러나 우리에게는 손톱깎이라고 하는 도구가 따로 있고 누구나 그 존재를 알고 있습니다. 중요한 점은 손톱깎이는 우리가 학교에서 누구한테 배워서 알거나 책에 그 사용법이 기술되어 있는 것이 아니라는 사실이다. 그냥 우리가 성인으로 커나가는 과정에서 자연스레 습득되는 상식일 뿐이다.

그렇다면 이러한 상식이 2022년 현재 인공지능 연구에서 왜 중요성이 있는 것일까? 이 질문에 대해서 몇 가지 사고실험을 해보자. 어떠한 마트를 관리하는 인공지능이 있다고 가정해 보다. 마트에서 판매하는 여러 상품은 당연하게도 외부적 변수에 의한 영향을 받게 된다. 대표적으로 비가 많이 내리는 장마철에는 막걸리의 판매량이 세 배 이상 증가한다. 마트 관리 인공지능에 상식적 추론 능력이 있다면 날씨와 상품 사이에 이러한 상관성이 있다는 것을 추론할 수 있어야 한다. 현재 밖에 비가 오는가? 그렇다면 현재 우리 매장에 막걸리 재고가 충분한가? 아니라면 막걸리를 바로 주문하여 채워 두어야 제대로 작동하는 '인공지능'이라고 할 수 있다. 막걸리와 비가 내리는 날씨와의 관계는 학습데이터에 명시적으로 기술되어 있지 않을 가능성이 더 크다. 하지만 적어도 한국 사람이라면 그리고 성인이라면 양자의 인과 관계를 경험적으로 알 것이다. 다른 예를 들어서 서빙 로봇을 생각해 보자. 서빙 로봇은 인공지능의 기능을 내장하여 음식을 주방에서 손님의 탁자까지 옮겨다 주는 일을 수행한다. 물론, 여기까지도 대단한 기술이다. 여러 장애물을 우회하여 동선을 실시간으로 계산하는 것이 쉬운 과제는 결코 아니다. 그러나 엄밀한 의미에서 인공지능 서빙이라고 하기에는 한계가 있다. 예컨대 우리가 식당에 갔다가 물을 엎지르는 일은 일상다반사로 자주 있는 일인데, 그럴 때 물을 엎지른 그 테이블로 냅킨을 스스로 갖다줄 수 있어야, 우리의 "상식"에 부합하는 "지능적" 행동이다. 인공지능의 이러한 기능까지 수행하는 것은 아직 기술적 장벽에 막혀있다.

인공지능 연구에서 최근 가장 중요한 이슈로 언급되는 "윤리"의 문제도 마찬가지이다. 앞서 서론에서 개괄한 바와 같이 상식은 많은 경우 윤리와 같은 맥락에서 논의된다. 즉, 못된 짓을 일삼는 사람에게 몰상식한 인간이라는 말을 하듯이, 결국 일상생활에서 윤리의 문제도 상식의 문제로 치환될 수 있다.

상식의 문제는, 현재의 인공지능 환경을 확률적 앵무새(stochastic parrot)로 비견한 관점에서 더 살펴볼 수 있다(Bender et al. 2021). 언어모델과 인간 언어의 가장 큰 차이점을 소통가능성(communicability)에서 찾는 것이 문제 제기의 시작이다. 인간은 언어를 사용하여 상호공통된 관심사를 전달하고 소통하려는 이른바 메타 의지가 있으며, 이러한 사실을 또한 서로 암묵적으로 알고 있다. 반면, 인공지능 언어모델은 이 의도를 하고 있다고 보기 어렵다. 이러한 의도가 배제된 상태에서 단순히 확률에 기반하여 단어의 조합을 결과값으로 반환한다는 점에서 언어 인공지능을 확률적 앵무새라고 비판하는 관점이다.

이 관점에서 상식 과제를 다시 이해해 볼 수 있다. 첫째, 상식 과제는 '자극의 빈곤(poverty of stimulus)'문제와 관련성을 가진다. 사람은 적은 수의 자극을 통해서도 특정

언어를 완전하게 '습득'할 수 있는 반면 인공지능은 방대한 양의 데이터를 반복적으로 살펴야만 제대로 '학습'을 수행할 수 있다. 즉, 학습데이터의 양과 질에 크게 의존하는 것이 인공지능의 특징이다. 상식은 명문화/성문화된 지식이 아니기 때문에, 도서 등으로 구성된 학습데이터에 '명시적으로' 기술되어 있지 않을 가능성이 더 크다. 그럼에도 불구하고, 특정 인공지능이 상식에 대한 판단을 준수하게 내릴 수 있다면, 그 인공지능은 단순하게 학습데이터의 겉면만을 학습한 것이 아니라고 유추할 수 있을 것이다. 둘째, 상식과제는 이른바 모라베크의 역설에 관계된다. 모라베크의 역설은 인간에게 매우 쉬운 것이 때로 컴퓨터에게는 매우 어렵고, 또 그 반대 관계도 성립할 수 있다는 것이다. 앞서 예를 든 바와 같이, 식당에서 물을 엎지르면 냅킨이 필요하다는 것은 유치원생들도 알 법한 아주 쉬운 추론이지만 인공지능에게는 쉽지 않다. 셋째, 확률적 앵무새라고 하는 비판은 언어 인공지능의 언어 사용이 아스퍼거 증후군 환자와 유사성을 보인다는 지적과 상통하는 바가 있다. 아스퍼거 증후군은 언어를 표면적으로만 해석하여 그 행간의 의미를 파악하지 못하는 현상을 말한다. 언어의 표면(surface language)를 넘어, 행간을 읽을 수 있는 능력을 갖추어야 진일보한 인공지능이라고 가늠할 수 있다. 현재의 언어 인공지능이 확률적 앵무새에 불과하다는 비판에서 벗어나려면, 행간의 의미, 문맥의 의미를 파악할 수 있어야 한다. 주지하다시피 상식은 대표적으로 행간에 숨겨져 있어 맥락적 이해를 요하는 지식체계이다. 이러한 이유에서 인공지능의 상식적 판단 능력이 평가의 대상이 되는 것이다.

3. 인공지능의 가추 추론 능력

3.1 인공지능과 가추 추론

이러한 상식 추론에 있어서 철학적 뼈대 역할을 하는 것이 흔히 말하는 가추 추론으로 영어로는 abductive reasoning이라고 한다. 가추 추론의 가장 대표적인 예시를 들자면 아래와 같다.

(13) ㄱ. If you hear hooves, think horse, not zebra.
ㄴ. 말발굽 소리를 들었다면,
얼룩말을 떠올리는 것이 아니라 그냥 말을 떠올려라!

(13ㄱ)은 실제로 미국 의과대학에서 수련의를 가르칠 때 사용하는 명제이다. 예컨대, 우리가 허리가 아파서 병원에 갔다고 가정하자. 사람이 허리가 아픈 이유는 아주 다양할 수 있고, 그중에는 심지어 신장에 문제가 있어서 허리가 아플 수도 있다. 물론 그 가능성은 상당히 낮다. 그런데 병원에 허리 통증 때문에 온 모든 환자에게 '당신이 허리가 아픈 이유는 신장 때문일 수도 있으니까 신장 검사도 합시다'라고 한다면 아마도 그 의사는 돌팔이 취급을 받을 것이다.

어떠한 현상 또는 사건이 발생했을 때, 그것을 가장 잘 설명할 것 같은 가설을 여러 개의 후보군 가운데 선택하는 것이 바로 가추 추론입니다. 가추 추론은 100여년전 퍼스

(Charles Sanders Peirce, 1839-1914)에 의해 체계를 갖추게 되었다. 퍼스는 언어학자, 기호학자, 철학자 등으로 알려져 있는데, 사실 처음 시작은 수학자이다. 이러한 관점에서 보면 자연세계 현상을 수학적으로 모델링하기 위해 출발한 논리체계가 바로 가추법이라고 할 수 있는데, 그 출발점이 현대 인공지능과 상당히 유사하다.

가추 추론을 압축적으로 정리하자면 바로 '설명적 가설'이다. 퍼스의 전기 관점에서 이러한 가설은 그 유명한 콩자루의 예시로 설명되었다. 아래에서 (14), (15), (16)은 차례로 연역법, 귀납법, 그리고 가추법의 예시이다.

(14) S1: 저 자루안의 콩은 모두 하얗다.
S2: 이 콩은 저 자루에서 나왔다.
S3: 그러므로 이 콩은 하얗다.

(15) S1: 이 콩들은 모두 저 자루에서 나왔다.
S2: 이 콩들은 모두 하얗다.
S3: 그러므로 저 자루안의 콩은 모두 하얗다.

(16) S1: 저 자루안의 콩은 모두 하얗다.
S2: 이 콩은 하얗다.
S3: 그러므로 이 콩은 저 자루에서 나왔다.

설명 가능성에 초점을 둔 가추법은 크게 사실, 현상, 그리고 설명의 축으로 도출되는데, 위 (16)에서 차례로 제시된 순서와 같다. 즉, 어떠한 사실을 바탕으로 주어진 현상을 가장 '그럴 듯 하게 설명'할 수 있는 명제를 만드는 것이다. 여기서 '그럴 듯 하다'라는 것은 다시 말해 정도성(degree)을 가진다는 것이고, 이러한 특성에서 학자에 따라 가추추론을 확률모델이라고 설명하기도 한다. 근본적으로 확률(likelyhood)에 기반한 현대 인공지능 언어모델과 궁합이 잘 맞는다.

약간 다른 관점에서 가추법을 세계에 대한 지식과 경험을 패턴화하는 도구라고 설명하기도 한다. 즉, 특정한 패턴을 인지하고 생성하는 방식으로 가정하는 것이다. 이러한 '패턴'의 차원에서 보아도 가추법은 자연언어에 대한 인공지능적 접근과 궤를 같이한다. 기술적 차원에서 우리가 현재 수행하고 있는 자연어처리의 과정을 "넓은 범위"의 패턴인식으로 본다면, 자연언어의 표현 안에 숨겨진 상식적 판단을 이끌어 내는 일은 충분히 가능하다. 언어학적 차원에서 우리가 현재 구축하고 있는 언어모델을 말뭉치에 기반한 파턴문법(pattern grammar, Hunston and Francis 2000; Busse and Moehlig-Falke 2019)의 일종이라고 가정한다면, 언어의 패턴, 맥락의 패턴, 문화의 패턴 등에 관한 가추 추론이 인공지능 연구에 들어오는 것은 지극히 합리적인 순서이다. 물론 사람과 컴퓨터는 메타의지 등등의 차원에서 차별화된다. 다만, 그 접근법에서 상당한 공통분모가 있다는 것이다.

3.2 최근 동향

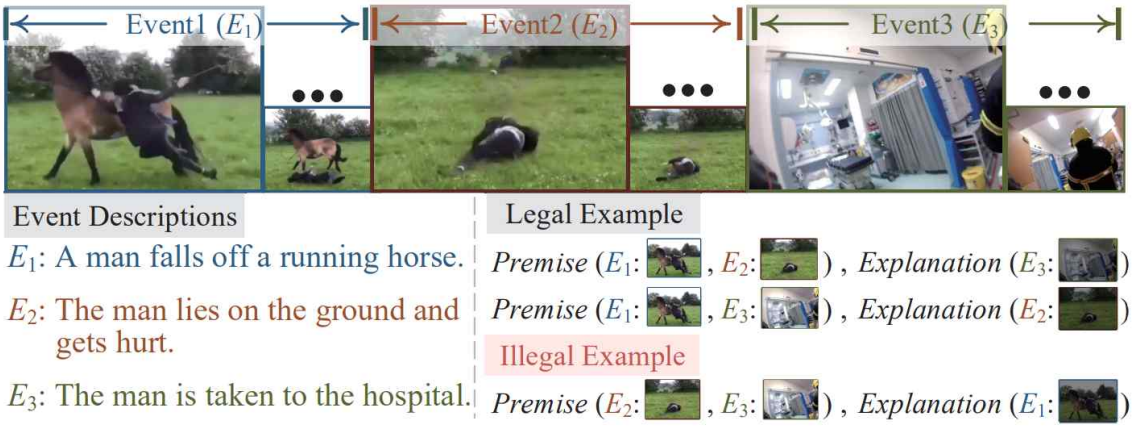
이러한 맥락에서 지난 몇 년간 여러 자연어처리 연구자들은 가추법에 기반한 상식 추론 능력을 인공지능에 이식하고 또 평가하는데 연구역량을 집중하여 왔다(Lin et al.

2019; Bhagavatula et al. 2019; Zellers et al. 2018; Huang et al. 2019; Bisk et al. 2020). 예컨대, (17)에서 아주 더운 여름 날(O1)인데, 그런 날에 기분이 좋아지려면 (O2) H− 혹은 H+ 가운데 어느 것이 더 설명적 타당성이 높을지 생각해보자.

(17) O1: It was a very hot summer day.
H−: He decided to run in the heat.
H+: He drank a glass of ice cold water.
O2: He felt much better!

물론 사람에 따라서 땡볕에서 달리기를 하면 기분이 좋아진다고 할 사람이 있을 수도 있다. 말발굽 소리가 들려서 밖에 나가보니 얼룩말일 수도 있는 것처럼 말이다. 서론에서 살핀 바와 같이 상식적 판단은 늘 오류가능성과 예외를 내포하기 마련이다. 그러나 그보다는 시원한 얼음물을 한잔 마셨을 때가 기분이 좋아질 가능성이 '상식적으로' 훨씬 크다. 이러한 상식적 판단을 인공지능도 사람처럼 할 수 있게끔 만드는 것이 현 단계 자연어처리의 핵심 과제이다.

2022년 최근에는 단순히 텍스트를 넘어, 아래에서 제시된 바와 같이 그림을 사용하는 가추추론 모델(Visual Abductive Reasoning)이 제시되었다(Liang et al. 2022). 이러한 흐름은 하나의 이야기 구조 안에서 추론을 도모한다는 점에서 특색이 있다. 이야기란 결국 언어, 맥락, 추론 등을 통괄하여 소통을 나누는 흐름이기 때문입니다. 그래서 퍼스의 관점에서 제대로 된 지능은 '이야기를 이해하고 만들 수 있는가' 더 나아가서 그림으로 그 이야기를 설명할 수 있는가의 문제로 확장된다. 즉, 그림까지 제대로 다룰 수 있어야 가추 추론이 완성되는 것이다. 이러한 관점에서 보면 VAR과 같은 접근은 연구의 당연한 발전 단계에 놓여 있다고 할 수 있다.



<그림 5> VAR의 예시 (Liang et al. 2022, 15567쪽)

4.

결론적으로 왜 인공지능에서 상식적 판단인가 혹은 언어모델에서 가추법인가에 대해서

정리를 하겠다. 가추 추론에 대한 보다 깊이 있는 논의는 이윤희(2013) 및 코블리(2022) 등을 참조하기 바란다.

무엇보다 가추법은 우리가 세상에 대한 지식을 습득하는 가장 첫번째 단추이기 때문이다. 우리가 일상생활 잘 인지하지는 못하지만 가장 널리 그리고 가장 기본적으로 사용되는 추론방식이 가추이다. 연역법과 귀납법에 전제되는 가설들은 추론에 앞서 선험적으로 존재하는 것이 아니므로, 현상을 탐구하고 추론하는 과정에서 가설의 형성 단계를 간과할 수 없기 때문이다. 이러한 관점에서 퍼스는 가추법을 세상에 대한 법칙화를 위해 선택된 그런 인간 진화의 결과물이라고 보았다(Fann, 1970). 즉, 우리의 뇌가 작동하는 기본 매커니즘이라는 것이다. 인공지능이 사람지능을 정말로 모델링한다라고 하면, 최소한 이러한 작동원리를 흉내 낼 수 있어야 할 것이다.

더 나아가서 인공지능의 일반화 가능성 그리고 확장 가능성과도 밀접하게 관련이 있는 것이 바로 상식 추론이다. 가추추론은 '그럴 듯 한 설명'을 선택하는 데 초점을 둔다. 따라서, 각각의 질문과 답변이 1:1로 코드화되어 있는 것이 아니라, 새로운 입력이 제시되었을 때 이에 적절하게 반응할 수 있는 새로운 출력을 생성하여 도출할 수 있어야 한다. 또한, 가추추론의 또 다른 의의는 비록 그 가설이 참이 아닐지라도 개연성 있게 (확률적으로) 설명했다는 데에 있다. 이러한 구성의 특성 자체가 인공지능의 파이프라인과 시너지 효과를 낼 수 있다.

신성일(2014), 《읽기 능력이 중학교 성적을 좌우한다》, 팜파스.
이윤희(2013), Semiotic Understanding of Mimetic Action from Peirce's Perspective: Towards Narrative Cognition and Symbolization, 《기호학 연구》 36, 197 217.
정도상(2017), 《모국어가 공부의 열쇠다 3단계》, 언어과학 .
코블리, 폴(2022), 《지적 대화를 위한 교양인의 기호학》, 팬덤북스 (이윤희 역).
Bender, Emily M., Timnit Gebru , Angelina McMillan Major, Shmargaret Shmitchell.(2021), On the Dangers of Stochastic Parrots : Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* Pages 610–623.
Bhagavatula, C., et al.(2019), Abductive commonsense reasoning. *arXiv preprint arXiv*:1908.05739.
Bisk, Y., Zellers, R., Gao, J., & Choi, Y.(2020), PIQA: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI conference on artificial intelligence*, pp. 7432–7439.
Brown, Tom, et al.(2020), Language models are few–shot learners. *Advances in neural information processing systems* 33, 1877–1901.
Busse, Beatrix, and Ruth Moehlig–Falke, eds.(2019), Patterns in Language and Linguistics: New Perspectives on a Ubiquitous Concept. Vol. 104. *Walter de Gruyter GmbH & Co KG*.
Engdahl, Elisabet & Enric Vallduví(1996), Information packaging in HPSG. *Edinburgh Working Papers in Cognitive Science* 12, 1–32.
Fann, Kuang Tih.(1970), *Peirce's theory of abduction*. Martinus Nijhoff.
Huang, L., Bras, R. L., Bhagavatula, C., & Choi, Y.(2019), Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. *arXiv preprint arXiv*:1909.00277.
Hunston, Susan, and Gill Francis.(2000), Pattern grammar: A corpus–driven approach to the lexical grammar of English. No. 4. *John Benjamins Publishing*.
Kojima, Takeshi, et al.(2022) Large Language Models are Zero–Shot Reasoners. *arXiv preprint arXiv*:2205.11916.
Liang, C., Wang, W., Zhou, T., & Yang, Y.(2022), Visual Abductive Reasoning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15565–15575.
Lin, Bill Yuchen, et al.(2019), CommonGen: A constrained text generation challenge for generative commonsense reasoning. *arXiv preprint arXiv*:1911.03705.
Rissman, Lilia, Kyle Rawlins, and Barbara Landau.(2015), Using instruments to understand

argument structure: Evidence for gradient representation. *Cognition* 142, 266-290.

Srivastava, Aarohi, et al.(2022) Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv*:2206.04615.

Wei, Jason, et al.(2021), Finetuned language models are zero-shot learners. *arXiv preprint arXiv*:2109.01652.

Zellers, R., Bisk, Y., Schwartz, R., & Choi, Y.(2018), SWAG: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv*:1808.05326.