한국어-영어 색채어 상대성에 따른 대형 언어 모델(LLM)의 행동 및 표상 편향 연구

권민준 송리안(멘토)

대전대신고등학교(Daejeon Daeshin High.) A.C.T.(KE)

ABSTRACT: 본 연구는 인간의 언어가 인지에 영향을 미친다는 언어 상대성 이론이, 영어 데이터 중심으로 학습된 대형 언어 모델(LLM)에서도 기계적 상대성으로 나타나는지 검증하고자 합니다. 러시아어의 파랑 구분 사례를 모델로 삼아, 특히 한국어와 영어 간의 색채 범주 차이에 초점을 맞춥니다. 연구팀은 고전적인 색채 식별 과제를 변형하여, 동일한 색상 자극에 대해 한국어와 영어 프롬프트가 모델의 선택 확률(H1)과 내부 표상 군집(H2)에 유의미한 차이를 유발하는지 측정합니다. 궁극적으로 이 연구는 LLM 의 언어 의존적 편향을 실증적으로 밝혀 AI 의 문화적, 언어적 정확성을 높이는 방안을 제시하는 데 기여할 것입니다.

I. 서론

언어가 인간의 사고를 결정한다는 사피어-워프 가설의 강한 형태는 오늘날 지지받지 않으나, 언어가 지각, 기억, 추론의 가중치를 바꾸고 특정 인지 전략을 유도한다는 약한 언어 상대성은 광범위하게 검증되었다. 가장 고전적이고 확실한 사례는 색채 지각 영역에서 발견된다. Winawer et al. (2007)은 러시아어 화자가 영어 화자와 달리 파란색을 siniy(진한 파랑)와 goluboy(연한 파랑)라는 두 개의 기본 범주로 나누며, 이 범주 경계에 걸친 색상들을 더 빠르게 구별한다는 것을 입증하였다. 이러한 인간의 인지 편향에 대한 논의는 최근 대형 언어 모델(LLM)의 등장으로 새로운 국면을 맞이하였다. 다국어 LLM 이 언어별 문법, 어순, 문화적 편향을 보이며, 심지어 입력 언어가 모델의 내부 인과 추론 경로를 다르게 프로그래밍할 수 있다는 기계적 상대성 연구가 보고되고 있다. 하지만 LLM 이 텍스트 토큰과 실제 세계의 물리적 실체(예: 색상)를 얼마나 잘 연결하는지에 대한 Color Grounding 연구는 대부분 영어 중심으로 이루어져, 비영어권 언어가 가진 미묘한 의미 범주를 모델이 어떻게 처리하는지에 대한 이해는 부족한 실정이다.

본 연구는 고전적 언어 상대성 연구의 가장 강력한 모델(러시아어 siniy/goluboy 사례)을 방법론적 준거로 삼아, LLM 의 영어 중심적 편향을 검증하고자 한다. 다만 러시아어-RGB 매핑 데이터셋 구축의 현실적 어려움을 고려하여, 본 연구는 한국어와 영어의 비교를 중심으로 이 방법론을 적용한다. 본 연구는 LLM 이 동일한 색채 구분 과제에서 프롬프트 언어(한국어 vs. 영어)에 따라 인간처럼 다른 행동 패턴(선택)을 보이는지, LLM 의 내부 임베딩 공간에서 동일한 RGB 값의 표상은 프롬프트 언어에 따라 다르게 군집화 되는지 가설을 세우고 실험 해보려고 한다.

II. 이론적 배경

2.1 언어의 상대성

언어 상대성은 언어가 지각, 기억, 추론의 가중치를 바꾸고 특정 인지 전략을 유도한다는 이론이다. 가장 고전적이고 확실한 사례는 색채 지각 영역에서 발견된다. Winawer et al. (2007)은 러시아어 화자가 영어 화자와 달리 파란색을 siniy(진한 파랑)와 goluboy(연한 파랑)라는 두 개의 기본 범주로 나누며, 이 범주 경계에 걸친 색상들을 더 빠르게 구별한다는 것을 입증하였다. 이는 언어 범주가 지각적 판단에 직접적인 영향을 미칠 수 있음을 시사한다.

2.2 기계적 상대성

기계적 상대성은 LLM 이 학습 데이터의 언어적 특성에 따라 편향된 추론이나 행동을 보이는 현상을 의미한다. 최근 연구들은 다국어 LLM 이 언어별 문법, 어순, 문화적 편향을 보이며, 심지어 입력 언어가 모델의 내부 인과 추론 경로를 다르게 프로그래밍할 수 있다고 보고하고 있다. 하지만 LLM 이 텍스트 토큰과 실제 세계의 물리적 실체(예: 색상)를 연결하는 Color Grounding 연구는 대부분 영어 중심으로 이루어져, 비영어권 언어의 고유한 범주를 모델이 어떻게 처리하는지에 대한 이해는 부족한 실정이다.

III. 한국어-영어 색채어 상대성 비교 분석

3.1 가설적 색채 범주 경계 및 자극 선정

본 연구는 새로운 인간 대상 실험을 수행하지 않는 대신, 기존 문헌 및 공인된 색채 표준 시스템을 분석하여 LLM 테스트에 사용될 가설적 범주 경계(Assumed Categorical Boundary)를 설정한다. 한국어의 초록색/연두색과 영어의 green/light green 의 용례를 분석한다. 특히 웹 컬러 표준(CSS/HTML), 먼셀(Munsell) 색 시스템, KSS 2821 등의 표준에서 green (#008000), light-green (#90EE90) 등의 기준점을 확보한다. 이 기준점들을 바탕으로, green 과 light green 사이의 스펙트럼을

그라데이션으로 생성한다. 이 스펙트럼에서 (a) 범주 내(Within-category) 자극(예: 둘 다 초록색 영역)과 (b) 범주 간(Cross-category) 자극(예: 하나는 초록색 영역, 하나는 연두색 영역)으로 명확히 구분되는 20 개의 최종 실험 자극 세트를 가설적으로 확정한다.

3.2 실험 프로토콜 설계

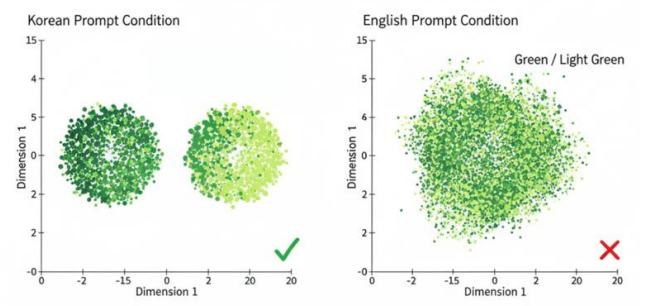
본 연구는 인간의 반응 시간(RT)을 LLM 의 선택 확률(Choice Probability)과 분류 일관성으로 치환하여 병렬적으로 설계한 LLM 병행평가 프로토콜을 사용한다.

• 대상 모델: GPT-4o, Gemini 1.5 Pro, Llama 3 등 다수의 다국어 모델 (영어 중심 모델과 다국어 특화 모델 포함)

• 자극(Stimuli): 가설적으로 확정된 20 개의 범주 내 및 범주 간 녹색 계열 RGB/Hex 코드 샘플.

3.3 LLM 행동 및 표상 분석

본 연구는 LLM 의 행동적, 표상적 편향을 검증하기 위해 두 단계의 분석을 수행합니다. 첫째, 행동 분석(H1 가설)을 위해 기준색(X)과 두 선택지(A, B)의 RGB 값을 범주 내 조건과 범주 간 조건으로 구성하여 LLM 에 제시하는 X-A-B 강제 선택 과제를 설계합니다. 이 과제는 동일한 RGB 자극에 대해 (a) 영어 프롬프트와 (b) 한국어 프롬프트로 각각 수백 회 반복 수행됩니다. 모델이 프롬프트를 처리한 후 선택지 A 와 B 토큰에 할당하는 최종 로짓(Logit) 값을 추출하며, 이는 모델의 원초적인 확신을 나타냅니다. 이 로짓 값은 소프트맥스(Softmax) 함수를 통해 A 를 선택할 확률과 B 를 선택할 확률이라는 표준화된 행동 지표로 변환됩니다. 수집된 데이터를 분석하기 위해 이항 로지스틱 회귀분석(Binomial Logistic Regression)을 실시합니다. 이 방법론을 채택한 이유는 종속 변수(범주 이탈 선택지 B 를 선택할 확률 등)가 0 또는 1 의 값을 갖는 이항적 특성을 가지기 때문이며, 이 모델은 독립 변수인 언어(한국어/영어)와 조건이 모델의 선택에 미치는 상호작용 효과를 통계적으로 직접 검증하는 데 가장 적합합니다. 만약 H1 가설이 참이라면, 이 상호작용 항은 통계적으로 유의미하게 관찰될 것입니다. 둘째, 이러한 행동적 차이가 모델의 내부 의미 표상(H2 가설)에서 기인하는지 확인하기 위해, 3.1 에서 확정된 20 개 자극의 RGB 값을 (a) 한국어 명명 프롬프트("이 색의 이름은?")와 (b) 영어 명명 프롬프트("What is the name of this color?")와 결합하여 모델에 입력합니다. 각 프롬프트 입력 시, 모델의 최종 레이어 은닉 상태를 해당 색상 자극의 내부 표상으로 추출합니다. 최종 레이어의 은닉 상태를 추출하는 이유는 이 벡터가 입력된 모든 정보(RGB 값과 언어적 맥락)를 종합하여 다음 단어를 예측하기 직전의, 가장 풍부하고 최종적인 의미 표상으로 간주되기 때문입니다.



추출된 임베딩을 t-SNE 을 사용해 2 차원으로 시각화

추출된 임베딩을 t-SNE 을 사용해 2 차원으로 시각화하였다. H2 가 참이라면, 한국어 조건에서만 초록색과 연두색의 명확한 군집(cluster) 분리가 나타난다. 이 시각화를 통해서 한국어로 LLM 에게 질문했을 때 LLM 은 연두색과 초록색을 완전히 다른 범주로 이해하고 있지만, 영어로 질문 했을 때는 LLM 은 Green 과 Light Green 을 엄격하게 나누지 않고, 그저 Light Green 을 Green 의 하위 종류로 인식한다. 이를 통해서 기계적 언어 상대성을 확인 할 수 있다.

IV. 시사 점 및 한계

5.1 연구의 기대 시사점

본 연구는 LLM 이 단순한 텍스트 패턴 학습을 넘어, 언어에 내재된 인지적 범주까지 학습(하거나 편향될) 수 있음을 실증적으로 보여준다. 이는 단순히 모델의 정확성을 높이는 것을 넘어, AI 의 문화적·언어적 정확성(Cultural Alignment)을 향상시키는 구체적인 엔지니어링 방안을 제시한다.

5.2 연구의 한계

본 연구는 비용 및 시간 제약으로 새로운 인간 대상 실험을 수행하지 않고, 기존 웹 표준 및 문헌을 기반으로 가설적 범주 경계를 사용하였다. 이는 실제 한국어 화자의 평균적인 지각 경계와 다소 차이가 있을 수 있으며, 연구 결과(H1, H2)의 타당성을 일부 제한할 수 있다. 본 연구는 색채 영역에 한정된다는 한계를 가진다. 또한 디지털 RGB 값으로 구현된 자극과 인간이 실제 물리적 환경에서 지각하는 색채 간의 차이가 존재할 수 있다.

VI. 결론

6.1 연구 결과 요약

본 연구는 인간의 고전적인 언어 상대성 효과(색채 지각)가 LLM 에서도 기계적 상대성으로 나타나는지를 한국어와 영어의 비교를 통해 검증하는 방법론을 제안하고, 그 기대 결과를 기술하였다. LLM 이 영어 중심의 색채 편향을 보일 것을 예측하며, 실험을

진행하였지만, 한국어로 질문 하였을 때 한국어의 언어적 특징에 따라서 군집도가 명확히 보이는 것으로 보였다.

6.2 향후 연구 방향 및 발전 전망

본 연구는 향후 공간(예: 끼다 와 put on/wear), 시간, 감정 등 다른 인지 영역으로 확장되어, LLM 의 보편성과 특수성을 탐구하는 후속 연구의 기반이 될 수 있다. 이는 비영어권 LLM 개발을 위한 문화 특화 데이터셋 구축의 필요성을 제언하는 근거가 될 것이다.

참고 문헌

Berlin, B., & Kay, P. (1969). Basic color terms: Their universality and evolution. University of California Press.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:2106.09685.

Kay, P., & Regier, T. (2006). Language, thought and color: Recent developments. Trends in Cognitive Sciences, 10(2), 51-54.

Liu, Z., et al. (2024). CultureLLM: Incorporating Cultural Differences into Large Language Models. arXiv preprint arXiv:2402.10946.

Masoud, M., Naous, T., & Baly, R. (2024). Investigating Cultural Alignment of Large Language Models. arXiv preprint arXiv:2402.13231.

Niedermann, J. P., et al. (2024). Studying the Effect of Globalization on Color Perception using Multilingual Online Recruitment and Large Language Models. arXiv preprint arXiv:2402.05739.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. Proceedings of the National Academy of Sciences, 104(19), 7780-7785.