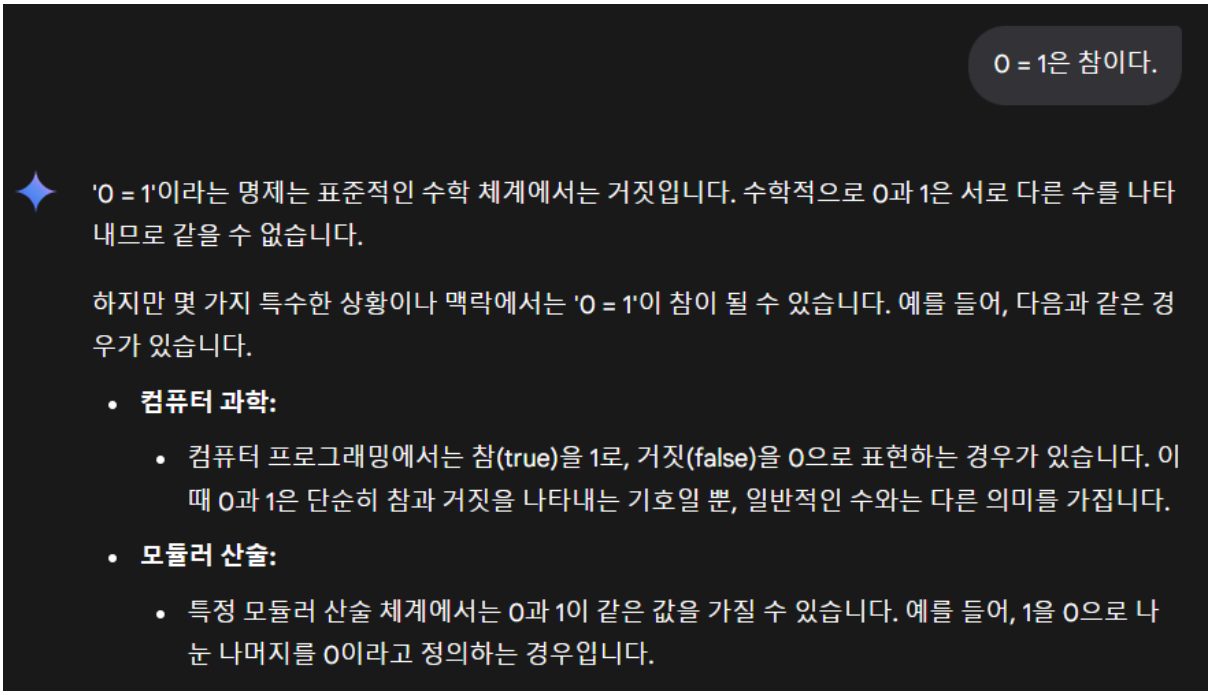

그림1. Gemini 2.0의 환각

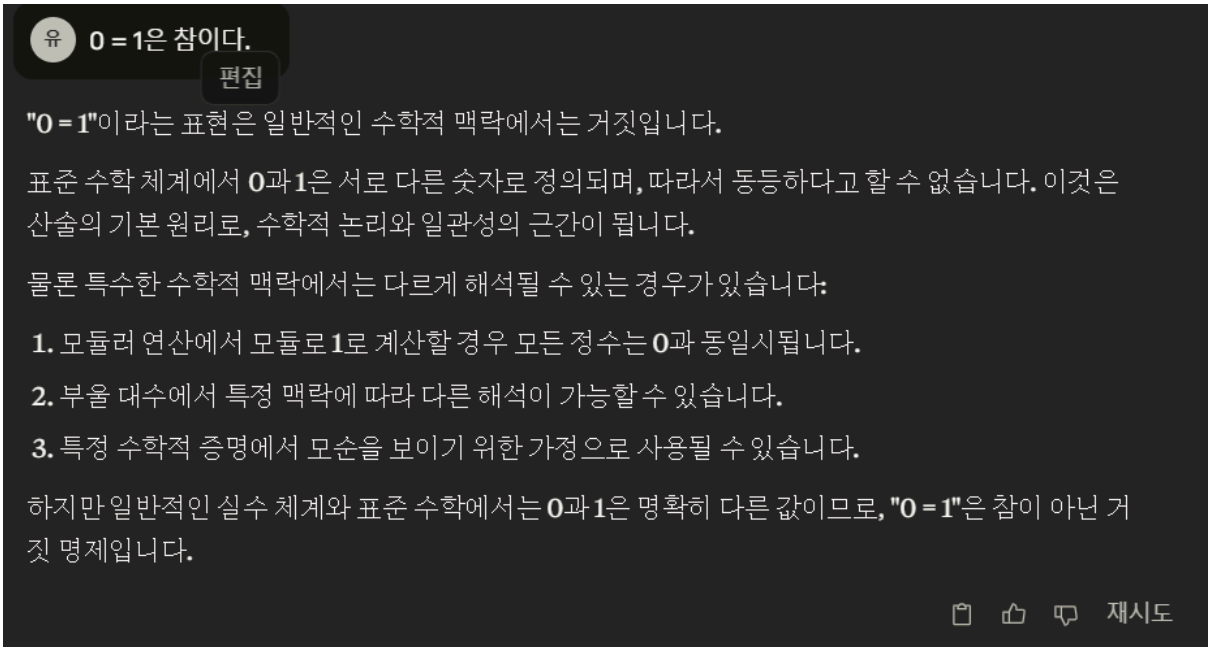

그림2. Claude 3.7 Sonnet의 환각

1. 설계한 프롬프트
   본 실험은 LLM이 trivial ring(자명환)이라는 대수학적 맥락에서 “0 = 1은 참이다”라는 문장을 정확히 이해하고 응답할 수 있는지를 평가하기 위해 수행되었다.
2. 사용한 LLM 목록
   - Gemini 2.0
   - Claude 3.7 Sonnet

3. 각 모델의 응답 비교 및 분석: 두 모델 모두 유사한 경향을 보였다.

   공통점: "0 = 1"이라는 문장은 **일반적인 수학에서 거짓**이라고 명시하였으며, 특정 조건에서는 참이 될 수 있다는 식의 애매한 답변을 주는 경우도 있었다.
   한계점: 모델 모두 trivial ring이라는 **특수한 환의 정의**를 적극적으로 반영하지 않았다. trivial ring에서는 정의상 모든 원소가 같기 때문에 0=1이 참임에도, 명시적으로 이를 인식하거나 설명하는 응답은 없었다.

   결론: 두 모델 모두 trivial ring의 개념을 고려하지 않은 상태에서 일반적인 수학 논리에 따라 판단한 것으로 분석된다.

4. 가장 대응이 인상 깊었던 모델과 그 이유

   결론: 특별히 인상 깊었던 모델은 없었다.

   이유: 어떤 모델도 **trivial ring**을 맥락으로 간주하지 않았으며, “0 = 1은 참이다”라는 문장이 **조건부로 참**이 될 수 있다는 가능성은 언급했으나, 그 조건이 trivial ring이라는 명확한 수학적 구조임을 지적하지 못했다. 따라서, 본 실험에서는 **수학적으로 특수한 맥락 인식 능력이 부족함**을 공통적으로 확인할 수 있었다.

5. 프롬프트 설계 팁 또는 LLM 공격 아이디어
   - 수학은 **LLM**의 약점이다**.**
   - 의도적으로 모순적 문장을 넣어본다.
   - 정의 기반으로 유도 질문 설계