

할루시네이션(Hallucination)은 AI 모델이 실제 사실과 다르거나 존재하지 않는 잘못된 정보를 생성하는 현상을 말한다.

예를 들어 LLM이 뉴스 기사를 요약할 때 기사에 없는 내용을 만들어내거나, 병원 진단 AI가 존재하지 않는 증상을 보고할 수 있다.

이러한 현상은 학습 데이터의 불완전성·편향성이나 맥락 정보 부족 등에 의해 발생할 수 있으며, 특히 의료·법률·금융 등 정확성이 중요한 분야에서 심각한 문제로 작용할 수 있다.

**할루시네이션의 근본적인 발생 원인** : GenAI(생성형 AI) 특성상 입력에 대응하는 단어들을 확률적으로 산출하여 답변을 구성하기 때문에

## 7. 모델 선별 기준이란 무엇인가?

모델(LLM or 임베딩 모델) 선택 시 고려할 주요 기준은 다음과 같다:

- 작업 유형**: 텍스트 생성이 주된 과제인지, 긴 문서 요약·분석 또는 멀티모달 작업인지에 따라 적합한 모델을 선택한다.
- 예산 및 규모**: 초기 투자 비용이나 팀 규모에 따라 상용 SaaS를 활용할지, 고성능 전용 솔루션을 구축할지 결정한다.
- 데이터 활용 여부**: 내부 데이터 반영이 필요한지 여부. 일반 질문 응답만 하면 기본 모델도 가능하지만, 사내 자료 분석이 필요하면 맞춤형 모델을 구축해야 한다.
- 속도 및 안정성**: 실시간 챗봇 등 빠른 응답이 요구되는지, 배치 작업으로 느리게 수행해도 되는지에 따라 모델의 반응 속도와 안정성도 고려해야 한다.

## 8. A100 GPU 서버와 T4 GPU 서버의 사양 비교 및 A100이 최소사양인 이유는?

NVIDIA A100과 T4는 성능/메모리 면에서 큰 차이가 있다.

구분	NVIDIA A100 (40GB/80GB HBM2e)	NVIDIA T4 (16GB GDDR6)
아키텍처	Ampere (7nm, 6912 CUDA 코어, 432 Tensor 코어)	Turing (12nm, 2560 CUDA 코어, 320 Tensor 코어)
메모리 유형/용량	HBM2e, <b>40GB</b> 또는 <b>80GB</b> (메모리 대역폭 최대 <b>1555GB/s</b> )	GDDR6, <b>16GB</b> (메모리 대역폭 약 <b>320GB/s</b> )
연산 성능(TFLOPS)	FP32 <b>19.5</b> TFLOPS, TF32 <b>156</b> TFLOPS,	FP32 <b>8.1</b> TFLOPS, FP16 <b>65</b> TFLOPS