Q. ChatGPT와 LLama 기반 Private LLM 성능 차이는 얼마나 나나?

A. GPT의 경우 성능을 따지는 구체적인 근거 수치(학습에 사용된 Parameter 수)는 공개되어 있지 않습니다.

다만, 우리가 시중에서 보는 상용 LLM의 경우, 특히 ChatGPT와 같은 B2C 서비스의 경우, LLM 모델로만 구현된 서비스가 아니라 내부적으론 파서부터 연산 함수 크롤러 등 엄청나게 많은 툴들이 서비스에 통합되어있습니다.

그래서 GPT 기반 모델과 LLM 성능지표(Context window, 모델 벤치마크)는 비슷한 Private LLM들이 사용자에게 느껴지는 체감 성능이 차이가 많이 나게 됩니다.

Q. 다양한 LLM 사용 중 어떤 LLM이 가장 좋다고 평가하는지?

A. 사용 목적(정밀도, 성능, 속도, 비용)에 따라 다르며, 여러 LLM을 실제 서비스에 연계해 평가합니다.
특히 상용LLM 은 범용적으로 사업화 하기에는 OpenAI GPT 가 가장 사용성은 좋다고 생각합니다. 하지만 점차 상향 평준화 되어가는 LLM 성능들에서 고객의 선호도와 비용, 특화 기능, 기존 인프라와의 호환성등을 고려하여 선택적으로 적용할 필요가 있습니다.

Q. 모델 선별 기준이란 무엇인가?

A. 모델(LLM or 임베딩 모델) 선택 시 고려할 주요 기준은 다음과 같습니다:

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

Q. ChatGPT가 아닌 Local LLM을 썼을 때 효과 중 환경 보안 강화의 의미는?

A. '환경 보안 강화'는 AI 솔루션이 구축되는 시스템 환경의 보안 수준을 높이는 것을 의미합니다.