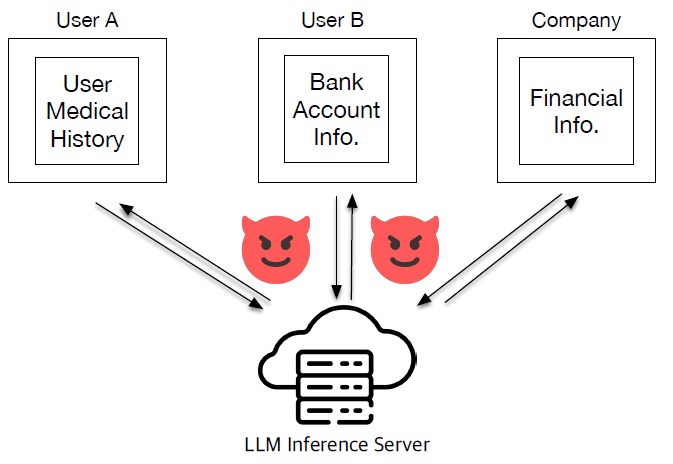
**Introduction**

2017년 “Attention Is All You Need”라는 논문 발표 이후 OpenAI, Google, 그리고 Meta와 같은 수 많은 기업들이 자신만의 거대언어모델(Large Language Model; LLM)을 개발해 출시해오고 있다. LLM의 추론 성능을 향상시키기 위해 다방면으로 수행된 연구들의 결과를 통해 확립된 사실은 추론 성능이 LLM의 모델 파라미터 개수와 비례한다는 점이다. 이에 따라 추론 성능을 향상시키기 위해 LLM의 크기는 나날이 커지고 있는 실정이다.

반면에 LLM의 크기가 커진다는 것은 연산장치 및 메모리와 같은 컴퓨팅 자원이 그만큼 더 요구된다는 점이다. 현재 LLM의 크기는 이미 상당히 커져 일반적으로 추론 요청을 하는 사용자의 로컬 컴퓨팅 머신에서는 효율적인 연산이 어렵거나 불가능한 탓에 컴퓨팅 자원이 풍부한 클라우드 서버에 추론 요청을 보내 연산한 뒤 결과물만 돌려받는 방식(Computation offloading scheme)을 채택한다. 이러한 방식을 통해 LLM 추론 서비스 사용자는 고성능 로컬 컴퓨팅 머신 없이도 서버의 컴퓨팅 자원을 사용해 쾌적한 추론 서비스를 누릴 수 있다.

하지만 위와 같은 Offloading scheme은 사용자의 민감정보를 포함하고 있는 추론 요청을 공격자가 존재할 수 있는 클라우드 서버에 어떠한 보호장치도 없이 보내기 때문에 보안적인 측면에서는 안전하지 않다. 아래 그림은 개인 사용자나 회사의 민감정보들을 포함한 LLM 추론 요청이 클라우드 서버로 내보내질때 공격자에 의한 민감정보 탈취 가능성을 보여준다.

**연구목표 및 필요성:** 정리하자면 매우 거대해진 LLM 추론 연산을 위해서 추론 요청을 클라우드 서버에 내보내 연산 후 결과를 되돌려 받는 오프로딩 방식은 효율적인 측면에서 필수적이다. 반면에 사용자 민감정보가 담긴 요청을 외부로 오프로딩 시키는 것은 공격자로 인한 탈취 및 위변조와 같은 공격으로부터 안전하지 않다. 이에 따라 본 연구의 목표는 추론 요청을 외부 가속기에 오프로딩 시켜 효율적인 추론 연산이 가능하게 하면서도 가능성있는 공격자로부터 사용자 민감정보를 안전하게 보호할 수 있는 보호 메커니즘 연구를 목표로 한다.

**Background**

**Trusted Execution Environment (TEE):** 본 연구가 궁극적으로 목표하는 것은 엄밀하게는 사용자 민감정보의 **기밀성(Confidentiality; 공격자로부터 민감정보의 의미를 보호)**을 보장하는 것이다. 본 연구에서의 한가지 가정은 외부 가속기는 공격으로부터 취약하고