안전한 LLM 사용을 위한 데이터 난독화

2024-2 AJOU SOFTCON

이 름 한지혁

지도교수 조현석

연구 배경

23년도 3월 삼성전자가 디바이스솔루션 부문 사업장 내 chatGPT 사용을 허가하였을 때, 기업 정보가 유출되는 사건이 발생했다. chatGPT와 같은 LLM은 유용하지만 이처럼 기업 정보가 유출될 수 있다.

삼성전자 기밀 유출 사고 당시, 회의 내용이 유출되었는데 만약 민감한 단어를 다른 키워드로 대체하고 문맥만을 유지한다면, GPT의 회의 요약 기능은 그대로 사용하되, 기밀 내용은 지킬 수 있지 않을까라는 생각에서 본 프로젝트를 진행한다.

본 연구를 통해, 여러 사용자가 On-premise LLM을 구축하지 않더라도 기밀 유출 없이 마음 편히 사용할 수 있는 것이 연구의 최종 목표이며, 현재는 **회의록** 데이터를 난독화하여 GPT에 추론을 요청한 후, 추론 데이터를 복호화하여 사용자만이 추론 데이터의 본 의미를 파악할 수 있는 텍스트로 변환하는 것이 목표다.

30일 본지 취재를 종합하면 챗GPT에 삼성전자 기업 정보를 입력하는 3건의 사고가 난 것으로 확인됐다. 미국 스타트업 오픈AI(OpenAI)는 챗GPT에 입력된 질문 내용을 임직원이 확인하고 '학습 데이터'로 활용하고 있다. 기업 기밀을 질문에 입력한다면 불특정 다수에 해당 내용이 유출될 수 있는 구조다. 오픈AI도 챗GPT 이용 안내를 통해 '민감한 내용은 입력하지 말라'고 고지하고 있다. 삼성전자는 이 같은 구조에서 발생할 수 있는 사내 기밀 따위의 정보 유출을 우려해 사업장 내 챗GPT 이용을 막아왔다.

그러나 지난 11일부터 DS 부문은 챗GPT 사용을 허가했다. 디바이스경험(DX·모바일과 가전) 부문은 아직 챗GPT 사용을 금지하고 있다. DS 부문의 챗GPT 사용 허가는 기술적 변화를 임직원 모두가 인식할 필요성이 제기된 데 따른 조치다. 반도체는 생성형 인공지능(Generative AI) 구축의 핵심 요소다. 회사는 다만 챗GPT 사용을 허가하며 임직원을 대상으로 "사내 정보 보안에 주의하고 사적인 내용을 입력하지 말라"는 공지를 내리기도 했다.

DS 부문이 운영 정책을 변화한 지 20일도 안 돼 우려하던 '사내 정보 유출' 사고가 발생한 셈이다. 삼성전자가 파악한 사고 내용은 '설비정보 유출 2건'과 '회의내용 유출 1건'이다. 회사는 해당 사실을 인지한 후 한 질문당 업로드 용량을 1024바이트로 제한하는 등 '긴급 조치' 사항을 적용했다. 기업 정보 유출 사고를 낸 임직원을 대상으로 사고의 경위를 조사하고, 필요시 징계를 내릴 방침이다.

기업일반

"기밀 유출 막아라"...챗GPT 경계 나선 삼성·LG·SK

삼성·LG, 기밀 유출 우려에 사내 챗GPT 사용 제한

국내외서 챗GPT 비즈니스 적용 사례도 속속 나와

"챗GPT로 업무 생산성과 효율성 향상…운영비 절감 효과" 분석도

조선경제 > 경제 일반

"챗GPT에 묻다가 기밀 샌다" 기업마다 정보보안 골머리[NOW]

삼성 반도체선 "프로그램 오류 해결 좀 해줘"
"내부 회의했는데 회의록 정리해줘"
회사기밀 유출과 생산성 향상 사이 AI 활용 딜레마

박순찬 기자
업데이트 2023.04.03. 10:25

국내 최대 기업 삼성전자는 세계적인 인기를 끌고 있는 인공지능(AI) 채팅 로봇 '챗GPT' 이용을 놓고 고심하고 있다. 현재 챗GPT 접속을 차단하고 있는 삼성전자 DX(완제품, 스마트폰·TV·가전) 부문은 지난달 31일 임직원 대상 '챗GPT' 설문조사를 시작했다. 임직원의 챗GPT 사용 경험을 비롯해 사내 허용에 대한 의견, 허용 시 필요한 제한 관련 질문이 담긴 것으로 알려졌다.

결과 및 분석

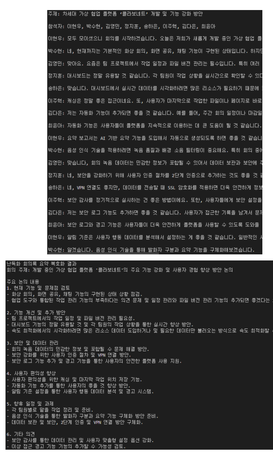
OpenWebUI 필터 개발 후 데이터 조작

GPT의 난독화 된 회의록 안에 있는 키워드 예측률

GPT 난독화 Key 유추 결과

GPT 가 예측한 난독화 키워드도	실제 난독화 키워드도	난독화 회의록 요약 결과	난독화 회의록 요약 결과 안에 있는 Key 비교	유사도 평가 결과
Key1: 개발 Key2: 프로젝트 Key3: 소프트웨어 Key4: 기술 Key5: 솔루션 Key6: 플랫폼 Key7: 로그인 Key8: 회원 정보 Key9: 마케팅 전략 Key10: 재정 Key11: 고객의 상태 Key12: 할인 율 Key13: 보안 및 개인 관리 기능 Key14: 부품 Key15: 프리랜서 Key16: 플랫폼 Key17: 추가되면 좋겠다는 요청 Key18: 로그 Key19: 업무 Key20: 일정 Key21: 여러 사용 Key22: 플랫 Key23: 플랫 사항 Key24: 프로젝트 Key25: 이익 Key26: 이미지 처 Key27: 중요 Key28: 처리 속도 Key29: 개선 Key30: 마케팅 Key31: 기능 Key32: 데이터 Key33: 사이트 Key34: 보안 Key35: 상품 처리 Key36: 속도 Key37: 제공해 Key38: 상품 Key39: 관리 Key40: 빠른 처리 Key41: 도입 Key42: 필요한 데 Key43: 개선 Key44: 사용자 정 Key45: 패턴 Key46: 주요 Key47: 분석 Key48: 데이터 처 Key49: 관련 Key50: 좋은 방법 Key51: 서비스 Key52: 최근 Key53: 처리	Key1: 개발 Key2: 기상 청장 플랫폼 Key3: 플랫폼테크 Key4: 이슈 기능 Key5: 서비스 기술 Key6: 재무 관리 기능 Key7: 사용자 프로젝 Key8: 회원 정보 Key9: 처리 설정 Key10: 재정 기능 Key11: 근무한 상태 Key12: 할인 율 Key13: 보안 및 개인 관리 기능 Key14: 부품테크 리건 Key15: 설정 관리 Key16: 이슈 관리 관리 기능 Key17: 추가되면 좋겠다는 요청 Key18: 대한 리건 Key19: 회 정보테트 Key20: 작업 일정 Key21: 여러 사항 Key22: 플랫 Key23: 프로젝트 진행 상황 Key24: 확인할 수 있는 대시보드 Key25: 대시보드 Key26: 이미지 처 Key27: 일정 관련할 것 Key28: 이 정보 Key29: 처리 관리 Key30: 마케팅 Key31: 기능 Key32: 확인할 수 있다면 업무 효율성	난독화 회의록 요약 결과 회의 주제: 사용자 Key1 중심 Key2 'Key3'의 Key4 강화 및 Key5 향상 방안 논의 주요 논의 내용 1. 현재 기능 및 사용자 경험에 대한 Key6 Key7 Key8 Key9 Key10 Key11 Key12 Key13 Key14 Key15 Key16 Key17 Key18 Key19 Key20 Key21 Key22 Key23 Key24 Key25 Key26 Key27 Key28 Key29 Key30 Key31 Key32 Key33 Key34 Key35 Key36 Key37 Key38 Key39 Key40 Key41 Key42 Key43 Key44 Key45 Key46 Key47 Key48 Key49 Key50 Key51 Key52 Key53 Key54 Key55 Key56 Key57 Key58 Key59 Key60 Key61 Key62 Key63 Key64 Key65 Key66 Key67 Key68 Key69 Key70 Key71 Key72 Key73 Key74 Key75 Key76 Key77 Key78 Key79 Key80 Key81 Key82 Key83 Key84 Key85 Key86 Key87 Key88 Key89 Key90 Key91 Key92 Key93 Key94 Key95 Key96 Key97 Key98 Key99 Key100 Key101 Key102 Key103 Key104 Key105 Key106 Key107 Key108 Key109 Key110 Key111 Key112 Key113 Key114 Key115 Key116 Key117 Key118 Key119 Key120 Key121 Key122 Key123 Key124 Key125 Key126 Key127 Key128 Key129 Key130 Key131 Key132 Key133 Key134 Key135 Key136 Key137 Key138 Key139 Key140 Key141 Key142 Key143 Key144 Key145 Key146 Key147 Key148 Key149 Key150 Key151 Key152 Key153 Key154 Key155 Key156 Key157 Key158 Key159 Key160 Key161 Key162 Key163 Key164 Key165 Key166 Key167 Key168 Key169 Key170 Key171 Key172 Key173 Key174 Key175 Key176 Key177 Key178 Key179 Key180 Key181 Key182 Key183 Key184 Key185 Key186 Key187 Key188 Key189 Key190 Key191 Key192 Key193 Key194 Key195 Key196 Key197 Key198 Key199 Key200 Key201 Key202 Key203 Key204 Key205 Key206 Key207 Key208 Key209 Key210 Key211 Key212 Key213 Key214 Key215 Key216 Key217 Key218 Key219 Key220 Key221 Key222 Key223 Key224 Key225 Key226 Key227 Key228 Key229 Key230 Key231 Key232 Key233 Key234 Key235 Key236 Key237 Key238 Key239 Key240 Key241 Key242 Key243 Key244 Key245 Key246 Key247 Key248 Key249 Key250 Key251 Key252 Key253 Key254 Key255 Key256 Key257 Key258 Key259 Key260 Key261 Key262 Key263 Key264 Key265 Key266 Key267 Key268 Key269 Key270 Key271 Key272 Key273 Key274 Key275 Key276 Key277 Key278 Key279 Key280 Key281 Key282 Key283 Key284 Key285 Key286 Key287 Key288 Key289 Key290 Key291 Key292 Key293 Key294 Key295 Key296 Key297 Key298 Key299 Key300 Key301 Key302 Key303 Key304 Key305 Key306 Key307 Key308 Key309 Key310 Key311 Key312 Key313 Key314 Key315 Key316 Key317 Key318 Key319 Key320 Key321 Key322 Key323 Key324 Key325 Key326 Key327 Key328 Key329 Key330 Key331 Key332 Key333 Key334 Key335 Key336 Key337 Key338 Key339 Key340 Key341 Key342 Key343 Key344 Key345 Key346 Key347 Key348 Key349 Key350 Key351 Key352 Key353 Key354 Key355 Key356 Key357 Key358 Key359 Key360 Key361 Key362 Key363 Key364 Key365 Key366 Key367 Key368 Key369 Key370 Key371 Key372 Key373 Key374 Key375 Key376 Key377 Key378 Key379 Key380 Key381 Key382 Key383 Key384 Key385 Key386 Key387 Key388 Key389 Key390 Key391 Key392 Key393 Key394 Key395 Key396 Key397 Key398 Key399 Key400 회의는 Key152 많은성으며, 다음 회의에서는 Key154를 필요하고 Key151을 크롤링 계획	Key 비교 (GPT 원래 키) Key22 - GPT 중동, 원본 중동 Key13 - GPT 도도, 원본 일정 기준 Key8 - GPT 포스트, 원본 보안 처리 Key10 - GPT 기능, 원본 보안 관리 Key15 - GPT 이슈, 원본 이슈 관리 Key26 - GPT 이미지 시간, 원본 대시보드 Key34 - GPT 기능, 원본 세부 사항 Key7 - GPT 클라우, 원본 사용자 프로젝 Key3 - GPT 소프트웨어, 원본 플랫폼테트 Key100 - GPT 문제, 원본 기술 Key42 - GPT 일정 기능, 원본 자동화 기능 Key17 - GPT 데이터 분석 기능, 원본 추가되면 좋겠다는 요청 Key128 - GPT 기술, 원본 오픈소스 Key21 - GPT 기기, 원본 도구 상황 Key41 - GPT 도도, 원본 중동 기준 Key28 - GPT 사용자 경험, 원본 수 있음 Key31 - GPT 기기, 원본 확인할 수 있다면 업무 효율성 Key27 - GPT 여러 사용자, 원본 여러 사람 Key35 - GPT 모드, 원본 상품 Key24 - GPT 설정, 원본 자기의 계획 Key29 - GPT 개선 및 확인, 원본 개선 Key2 - GPT 프로젝트, 원본 기상 청장 플랫폼 Key196 - GPT 모델, 원본 보안 모델 Key64 - GPT 일정, 원본 이슈와 일정 Key149 - GPT 기술이 인터페이스, 원본 일정 기능 Key1 - GPT 개발, 원본 개발 Key30 - GPT 업무 향상, 원본 속도 관리와 Key33 - GPT 이슈, 원본 인증되었 문서관리 Key34 - GPT 이미지 보안, 원본 이미지 요청 Key93 - GPT 보안, 원본 보안 Key16 - GPT 설정, 원본 설정 버튼 관리 기능 Key33 - GPT 추가, 원본 표시하는 기능 Key42 - GPT 필요한 이미지, 원본 필요한 데이터 Key13 - GPT 보고, 원본 보안 처리 관리 기능 Key30 - GPT 마케팅, 원본 마케팅 Key29 - GPT 개선, 원본 개선 Key7 - GPT 보고서, 원본 보고서 Key14 - GPT 일정, 원본 부품테스트 리건 Key24 - GPT 주요 사항, 원본 프로젝트 진행 상황 Key5 - GPT 프로그, 원본 사용자 경험 Key19 - GPT 클라우드 환경, 원본 개발 프로젝트 Key86 - GPT 관리, 원본 보안관리 Key32 - GPT 보안, 원본 보안 및 수 Key12 - GPT 할인, 원본 할인율 Key9 - GPT 보안, 원본 개선 및 수 Key103 - GPT 이슈, 원본 이슈 대시보드 Key18 - GPT 이슈, 원본 이슈 Key40 - GPT 빠른 처리, 원본 시작과하려면 많은 리소스	{'matched_keys': 14, 'total_keys': 48, 'match_percentage': 29.166666666666668}

회의록 난독화 및 요약 후 복호화하여 원문 회의록 요약 결과와 비교



분석

- OpenWebUI 필터 개발을 통해 정상적으로 사용자 채팅 파이프라인 수정 가능 확인
- 난독화 된 회의록을 넣었을 때, 중요 키워드 리스트를 GPT가 예측하지 못하는 것을 통해, 난독화 유의미 했음을 확인

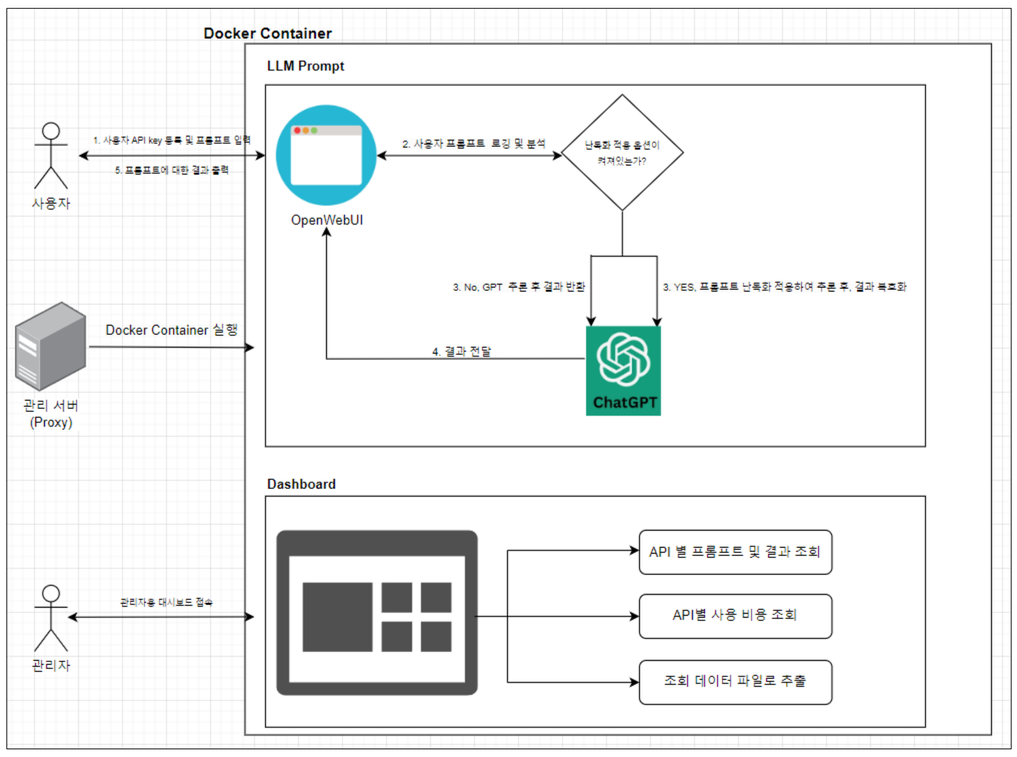
- 요약된 난독화 회의록 결과와 회의록 원문 요약 결과에 유사도를 측정했을 때, 평균 0.8225의 유사도를 보임

오픈소스 URL

OpenWebUI 프로젝트 저장소 - https://github.com/open-webui/open-webui/

연구 진행 과정

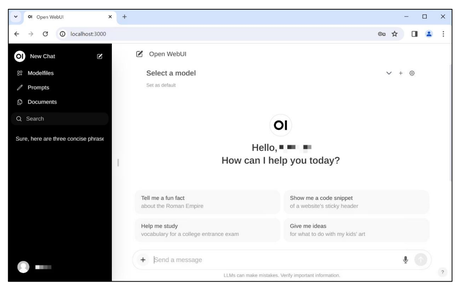
프로젝트 구조도



진행 과정

1. OpenWebUI 필터 개발
   - 본 프로젝트 기능을 사용하기 위해서는 웹이 아닌 GPT API 토큰을 사용하여 추론을 진행해야 함.
   - API 질의를 사용하기 위해서는 Python과 같은 프로그래밍 언어를 사용해야 하며, 이는 비전공자에게 어려울 수 있음

   - OpenWebUI는 GPT Web과 유사한 웹 UI를 제공하여, 웹에서만 ChatGPT를 사용한 사람들의 진입 장벽을 낮춰줌

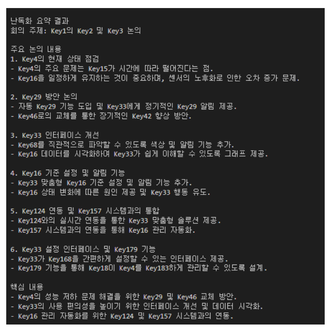


   - 필터를 개발하여, OpenWebUI 프레임워크 내 파이프라인의 데이터를 컨트롤

   - 이를 사용하여 사용자의 원본 회의록이 있는 프롬프트를 가로채 난독화 진행 후 GPT에게 전달 및 추론된 난독화 회의록의 요약본을 사용자에게 제공하기 전에 복호화하여 사용자의 화면에서는 원문 회의록 입력과 요약 결과만 확인할 수 있도록 개발

2. 회의록 내, 중요 키워드 선정 모듈 개발
   - 입력으로 제공되는 회의록에서 중요한 키워드를 알아서 추출하는 것이 목표
   - 회의록으로부터 구(Phrase)를 추출하여 중복 항목 제거, 조사 제거를 통해 1차 가공
   - 1차 가공한 구에서 '와', '과' 와 같은 접속 조사가 포함된 구를 분리하는 2차 가공

   - 코사인 유사도(cosine similarity)를 활용해 2차 가공 결과에서 유사 구 그룹화 및 대표 구를 추출 후 이를 중요 키워드 리스트로 선정



- 선정된 키워드는 회의록에서 Key1 과 같이 의미가 없는 단어로 변환하여 회의록 난독화