

Spis treści

Wprowadzenie	11
1. Co poszło nie tak z chatbotami?	17
Pomówmy o projekcie Tay	17
Gwałtowny upadek Tay	18
Dlaczego doszło do afery z Tay?	19
To trudny problem	21
2. Lista OWASP top 10 dla aplikacji używających dużych modeli językowych	23
Fundacja OWASP	24
Lista top 10 projektu aplikacji wspomaganych przez duże modele językowe	25
Realizacja projektu	25
Przyjęcie	26
Klucz do sukcesu	27
Ta książka i lista top 10	28
3. Architektury i granice zaufania	29
Sztuczna inteligencja, sieci neuronowe i duże modele językowe — czym się różnią?	29
Rewolucja transformerów — źródło, wpływ i powiązanie z dużymi modelami językowymi	31
Pochodzenie transformerów	31
Wpływ architektury transformerów na sztuczną inteligencję	32
Rodzaje aplikacji wspomaganych przez duże modele językowe	33
Architektura aplikacji wspomaganych przez duże modele językowe	35
Granice zaufania	36
Model	38
Interakcja z użytkownikiem	39
Zbiór danych użytych do wytrenowania modelu	40
Dostęp do bieżących zewnętrznych źródeł danych	41
Dostęp do usług wewnętrznych	43
Podsumowanie	43

4. Wstrzykiwanie promptu	44
Przykłady ataków typu wstrzykiwanie promptu	45
Natarczywa sugestia	45
Psychologia odwrotna	46
Wprowadzenie w błąd	47
Uniwersalne i zautomatyzowane prompty antagonistyczne	48
Wpływ ataków polegających na wstrzykiwaniu promptu	49
Bezpośredni i pośredni atak polegający na wstrzykiwaniu promptu	50
Bezpośrednie wstrzykiwanie promptów	50
Pośrednie wstrzykiwanie promptów	51
Najważniejsze różnice	51
Łagodzenie skutków ataku polegającego na wstrzykiwaniu promptu	52
Ograniczanie częstotliwości wykonywania zapytań do modelu	53
Filtrowanie danych wejściowych za pomocą reguł	53
Filtrowanie za pomocą dużego modelu językowego specjalnego przeznaczenia	54
Dodawanie struktury promptu	55
Trenowanie antagonistyczne	57
Definicja pesymistycznych granic zaufania	58
Podsumowanie	59
5. Czy duży model językowy może wiedzieć zbyt wiele?	60
Rzeczywiste przykłady	61
Lee Luda	61
GitHub Copilot i OpenAI Codex	62
Metody zdobywania wiedzy	64
Trenowanie modelu	65
Trenowanie modelu podstawowego	65
Kwestie dotyczące bezpieczeństwa modeli podstawowych	66
Dostrajanie modelu	67
Niebezpieczeństwo związane z trenowaniem	67
Technika RAG	69
Bezpośredni dostęp do sieci WWW	70
Uzyskiwanie dostępu do bazy danych	74
Uczenie się na podstawie interakcji z użytkownikiem	79
Podsumowanie	81
6. Czy modele językowe śnią o wirtualnych baranach?	83
Dlaczego duży model językowy ulega halucynacji?	84
Rodzaje halucynacji	85
Przykłady	85
Nieistniejące precedensy prawne	86
Proces dotyczący chatbota linii lotniczej	87

Nieumyślne skrzywdzenie człowieka	89
Halucynacje pakietu otwartoźródłowego	90
Kto ponosi odpowiedzialność?	91
Najlepsze praktyki w zakresie zmniejszania niebezpieczeństwa	93
Rozszerzenie wiedzy ściśle związanej z daną dziedziną	93
Łańcuch myśli, który zachęca do większej dokładności	95
Mechanizmy przekazywania informacji zwrotnych — potężne możliwości	96
danych wejściowych użytkownika w zmniejszeniu niebezpieczeństwa	97
Przejrzysta komunikacja dotycząca oczekiwanego sposobu użycia modelu i jego ograniczeń	99
Szkolenie użytkowników — wzmacnianie ich dzięki wiedzy	101
Podsumowanie	102
7. Nie ufaj nikomu	102
Wyjaśnienie zerowego zaufania	103
Skąd ta paranoja?	104
Implementacja architektury zerowego zaufania dla dużego modelu językowego	105
Obserwacja pod kątem nadmiernej działalności	106
Zabezpieczanie obsługi danych wyjściowych	109
Tworzenie własnego filtra danych wyjściowych	112
Wyszukiwanie danych osobowych za pomocą wyrażenia regularnego	112
Sprawdzanie pod kątem toksyczności	113
Połączenie filtrów z dużym modelem językowym	114
Oczyszczanie w celu zapewnienia bezpieczeństwa	115
Podsumowanie	116
8. Nie trać głowy	117
Ataki typu DoS	118
Atak ilościowy	118
Atak na protokół	119
Atak na warstwę aplikacji	120
Epicki atak typu DoS na firmę Dyn	120
Przygotowanie ataku typu DoS na duży model językowy	121
Ataki na ograniczone zasoby	121
Wykorzystanie okna kontekstu	122
Nieprzewidywalne dane wejściowe użytkownika	123
Ataki typu DoW	124
Klonowanie modelu	126
Strategie łagodzenia skutków ataku	126
Mechanizmy obronne ściśle związane z daną dziedziną	126
Weryfikacja danych wejściowych i ich oczyszczanie	127
Niezawodne ograniczanie częstotliwości wykonywania zapytań	127

Ograniczanie poziomu użycia zasobów	128
Monitorowanie i ostrzeganie	128
Finansowe wartości progowe i ostrzeżenia	128
Podsumowanie	129
9. Znajdź najbliższe ogniwo	130
Podstawy łańcucha dostaw	131
Bezpieczeństwo łańcucha dostaw oprogramowania	132
Incydent związany z agencją Equifax	132
Incydent związany z SolarWinds	133
Luka w zabezpieczeniach Log4Shell	135
Poznanie łańcucha dostaw w przypadku dużego modelu językowego	136
Ryzyko w modelu otwartoźródłowym	137
Zatrucie zbioru danych użytych do wytrenowania modelu	138
Przypadkowo niebezpieczne dane uczące	139
Niebezpieczne wtyczki	139
Tworzenie artefaktów przeznaczonych do śledzenia łańcucha dostaw	140
Waga SBOM	140
Karty modeli	141
Karta modelu a zestawienie komponentów oprogramowania	142
CycloneDX — standard zestawienia komponentów oprogramowania	144
Powstanie ML-BOM	145
Utworzenie przykładowego zestawienia ML-BOM	146
Przyszłość bezpieczeństwa łańcucha dostaw dużego modelu językowego	149
Podpis cyfrowy i znak wodny	149
Bazy danych i klasyfikacje luk w zabezpieczeniach	150
Podsumowanie	155
10. Wyciąganie wniosków na przyszłość	156
Przegląd listy OWASP top 10 dla aplikacji wspomaganych przez duże modele językowe	156
Studia przypadków	157
Dzień Niepodległości — głośna katastrofa	158
2001: Odyseja kosmiczna	161
Podsumowanie	164
11. Zaufaj procesowi	165
Ewolucja ruchu DevSecOps	166
MLOps	167
LLMOps	167
Wbudowanie bezpieczeństwa do procesu LLMOps	168

Bezpieczeństwo w trakcie procesu programistycznego związanego z dużym modelem językowym	169
Zabezpieczanie potoku CI/CD	169
Ściśle związane z dużym modelem językowym narzędzia sprawdzania stanu bezpieczeństwa	170
Zarządzanie łańcuchem dostaw	172
Chroń aplikację za pomocą mechanizmów obronnych	173
Rola mechanizmu obronnego w strategii bezpieczeństwa dużego modelu językowego	174
Otwartoźródłowe i komercyjne rozwiązania w zakresie mechanizmów obronnych	175
Łączenie niestandardowych i gotowych mechanizmów obronnych	176
Monitorowanie aplikacji	176
Rejestrowanie każdego promptu i odpowiedzi	176
Scentralizowane zarządzanie zdarzeniami i dziennikami zdarzeń	177
Analiza sposobu działania użytkownika i encji	177
Utworzenie własnego czerwonego zespołu dla sztucznej inteligencji	177
Zalety czerwonego zespołu sztucznej inteligencji	179
Czerwone zespoły kontra pentesterzy	180
Narzędzia i podejścia	181
Nieustanne usprawnianie	182
Tworzenie i dostrajanie mechanizmów obronnych	182
Zarządzanie jakością i dostępem do danych	183
Użycie techniki RLHF	183
Podsumowanie	185

12. Praktyczny framework zapewnienia bezpieczeństwa odpowiedzialnej sztucznej inteligencji 186

Moc	187
Procesory graficzne	188
Chmura	189
Ruch otwartoźródłowy	190
Wielomodalność	192
Autonomiczne agenty	193
Odpowiedzialność	194
Framework RAISE	195
Lista rzeczy do sprawdzenia za pomocą frameworka RAISE	201
Podsumowanie	202