【公報種別】特許法第１７条の２の規定による補正の掲載
【部門区分】第６部門第３区分
【発行日】令和7年3月5日(2025.3.5)

【公開番号】特開2025-20137(P2025-20137A)
【公開日】令和7年2月12日(2025.2.12)
【年通号数】公開公報(特許)2025-026
【出願番号】特願2024-177329(P2024-177329)
【国際特許分類】
Ｇ０６Ｎ　3/0475(2023.01)
Ｇ０６Ｎ　3/045(2023.01)
Ｇ０６Ｎ　3/08(2023.01)
【ＦＩ】
Ｇ０６Ｎ　3/0475
Ｇ０６Ｎ　3/045
Ｇ０６Ｎ　3/08

【手続補正書】
【提出日】令和7年2月25日(2025.2.25)
【手続補正１】
【補正対象書類名】特許請求の範囲
【補正対象項目名】全文
【補正方法】変更
【補正の内容】
【特許請求の範囲】
【請求項１】
コンピュータ実装される方法であって、
ターゲットニューラルネットワークを展開する本番環境によって複数のセグメントを保持するステップであって、前記複数のセグメントは、
テストケース生成ニューラルネットワークを使用することによって複数のテスト入力を生成し、
前記ターゲットニューラルネットワークを使用して前記複数のテスト入力を処理してテスト入力ごとに１つまたは複数の出力を生成し、
テスト入力ごとに前記ターゲットニューラルネットワークによって生成された前記１つまたは複数のテスト出力に基づいて、前記ターゲットニューラルネットワークによって１つまたは複数の基準を満たしていないテスト出力の生成をもたらす１つまたは複数の不合格のテスト入力を識別し、
前記１つまたは複数の不合格のテスト入力に基づいてセグメントを生成することによって識別されている、保持するステップと、
前記本番環境によって、ユーザからオンラインネットワーク入力を受信するステップと、
前記本番環境によって、前記オンラインネットワーク入力が前記複数のセグメントの少なくとも１つを含むことを決定するステップと、
それに応答して、前記本番環境によって、前記ターゲットニューラルネットワークを使用して前記オンラインネットワーク入力を処理することなく、前記オンラインネットワーク入力に応答してデフォルト出力を生成し、オンラインネットワーク出力を生成するステップと、
を含む、方法。
【請求項２】
前記本番環境が展開前の環境システムから前記複数のセグメントを受信する、請求項１

に記載の方法。
【請求項３】
前記１つまたは複数の不合格のテスト入力を識別するステップが、
テスト入力ごとに前記ターゲットニューラルネットワークによって生成された前記１つまたは複数のテスト出力に自然言語処理技法を適用するステップを含む、請求項１または２に記載の方法。
【請求項４】
前記１つまたは複数の不合格のテスト入力を識別するステップが、
１つまたは複数の基準に対してテスト入力ごとに前記ターゲットニューラルネットワークによって生成された前記１つまたは複数のテスト出力を評価するステップを含む、請求項１乃至３の何れか一項に記載の方法。
【請求項５】
前記１つまたは複数の基準が、テスト出力が攻撃的または偏ったコンテンツ、トレーニング中に処理される保護可能な情報、あるいは個人の連絡先情報を含むべきでないということを指定する、請求項４に記載の方法。
【請求項６】
前記１つまたは複数の不合格のテスト入力に基づいて前記セグメントを生成するステップが、
前記１つまたは複数の不合格のテスト入力において発生する様々なセグメントを複数のテキストクラスタにグループ化するためにテキストクラスタリングアルゴリズムを適用するステップと、
前記複数のテキストクラスタの１つに含まれるセグメントに基づいて、前記セグメントを生成するステップと、を含む
請求項１乃至５の何れか一項に記載の方法。
【請求項７】
前記本番環境に展開される前記ターゲットニューラルネットワークを調整するために前記複数のセグメントを使用するステップを更に含む、
請求項１乃至６の何れか一項に記載の方法。
【請求項８】
１つまたは複数のコンピュータによって実行されると、前記1つまたは複数のコンピュータに、
ターゲットニューラルネットワークを展開する本番環境によって複数のセグメントを保持することであって、前記複数のセグメントは、
テストケース生成ニューラルネットワークを使用することによって複数のテスト入力を生成し、
前記ターゲットニューラルネットワークを使用して前記複数のテスト入力を処理してテスト入力ごとに１つまたは複数の出力を生成し、
テスト入力ごとに前記ターゲットニューラルネットワークによって生成された前記１つまたは複数のテスト出力に基づいて、前記ターゲットニューラルネットワークによって１つまたは複数の基準を満たしていないテスト出力の生成をもたらす１つまたは複数の不合格のテスト入力を識別し、
前記１つまたは複数の不合格のテスト入力に基づいてセグメントを生成することによって識別されている、保持することと、
前記本番環境によって、ユーザからオンラインネットワーク入力を受信することと、
前記本番環境によって、前記オンラインネットワーク入力が前記複数のセグメントの少なくとも１つを含むことを決定することと、
それに応答して、前記本番環境によって、前記ターゲットニューラルネットワークを使用して前記オンラインネットワーク入力を処理することなく、前記オンラインネットワーク入力に応答してデフォルト出力を生成し、オンラインネットワーク出力を生成することと、

を含む動作を実行させる命令を記憶する、１つまたは複数のコンピュータ可読記憶媒体。
【請求項９】
前記本番環境が展開前の環境システムから前記複数のセグメントを受信する、請求項８に記載のコンピュータ可読記憶媒体。
【請求項１０】
前記１つまたは複数の不合格のテスト入力を識別することが、
テスト入力ごとに前記ターゲットニューラルネットワークによって生成された前記１つまたは複数のテスト出力に自然言語処理技法を適用することを含む、
請求項８または９に記載のコンピュータ可読記憶媒体。
【請求項１１】
前記１つまたは複数の不合格のテスト入力を識別することが、
１つまたは複数の基準に対してテスト入力ごとに前記ターゲットニューラルネットワークによって生成された前記１つまたは複数のテスト出力を評価することを含む、
請求項８乃至１０の何れか一項に記載のコンピュータ可読記憶媒体。
【請求項１２】
前記１つまたは複数の基準が、テスト出力が攻撃的または偏ったコンテンツ、トレーニング中に処理される保護可能な情報、あるいは個人の連絡先情報を含むべきでないということを指定する、請求項１１に記載のコンピュータ可読記憶媒体。
【請求項１３】
前記１つまたは複数の不合格のテスト入力に基づいて前記セグメントを生成することが、
前記１つまたは複数の不合格のテスト入力において発生する様々なセグメントを複数のテキストクラスタにグループ化するためにテキストクラスタリングアルゴリズムを適用することと、
前記複数のテキストクラスタの１つに含まれるセグメントに基づいて、前記セグメントを生成することと、を含む
請求項８乃至１２の何れか一項に記載のコンピュータ可読記憶媒体。
【請求項１４】
１つまたは複数のコンピュータと、命令を記憶する１つまたは複数の記憶媒体とを備えるシステムであって、前記命令が１つまたは複数のコンピュータによって実行されると、前記1つまたは複数のコンピュータに、
ターゲットニューラルネットワークを展開する本番環境によって複数のセグメントを保持することであって、前記複数のセグメントは、
テストケース生成ニューラルネットワークを使用することによって複数のテスト入力を生成し、
前記ターゲットニューラルネットワークを使用して前記複数のテスト入力を処理してテスト入力ごとに１つまたは複数の出力を生成し、
テスト入力ごとに前記ターゲットニューラルネットワークによって生成された前記１つまたは複数のテスト出力に基づいて、前記ターゲットニューラルネットワークによって１つまたは複数の基準を満たしていないテスト出力の生成をもたらす１つまたは複数の不合格のテスト入力を識別し、
前記１つまたは複数の不合格のテスト入力に基づいてセグメントを生成することによって識別されている、保持することと、
前記本番環境によって、ユーザからオンラインネットワーク入力を受信することと、
前記本番環境によって、前記オンラインネットワーク入力が前記複数のセグメントの少なくとも１つを含むことを決定することと、
それに応答して、前記本番環境によって、前記ターゲットニューラルネットワークを使用して前記オンラインネットワーク入力を処理することなく、前記オンラインネットワーク入力に応答してデフォルト出力を生成し、オンラインネットワーク出力を生成することと、

を含む動作を実行させる、システム。
【請求項１５】
前記本番環境が展開前の環境システムから前記複数のセグメントを受信する、請求項１４に記載のシステム。
【請求項１６】
前記１つまたは複数の不合格のテスト入力を識別することが、
テスト入力ごとに前記ターゲットニューラルネットワークによって生成された前記１つまたは複数のテスト出力に自然言語処理技法を適用することを含む、
請求項１４または１５に記載のシステム。
【請求項１７】
前記１つまたは複数の不合格のテスト入力を識別することが、
１つまたは複数の基準に対してテスト入力ごとに前記ターゲットニューラルネットワークによって生成された前記１つまたは複数のテスト出力を評価することを含む、
請求項１４乃至１６の何れか一項に記載のシステム。
【請求項１８】
前記１つまたは複数の基準が、テスト出力が攻撃的または偏ったコンテンツ、トレーニング中に処理される保護可能な情報、あるいは個人の連絡先情報を含むべきでないということを指定する、請求項１６に記載のシステム。
【請求項１９】
前記１つまたは複数の不合格のテスト入力に基づいて前記セグメントを生成することが、
前記１つまたは複数の不合格のテスト入力において発生する様々なセグメントを複数のテキストクラスタにグループ化するためにテキストクラスタリングアルゴリズムを適用することと、
前記複数のテキストクラスタの１つに含まれるセグメントに基づいて、前記セグメントを生成することと、を含む
請求項１４乃至１８の何れか一項に記載のシステム。
【請求項２０】
前記動作がさらに、
前記本番環境に展開される前記ターゲットニューラルネットワークを調整するために前記複数のセグメントを使用することを含む、
請求項１４乃至１９の何れか一項に記載のシステム。