

(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2025年1月9日(09.01.2025)



(10) 国際公開番号

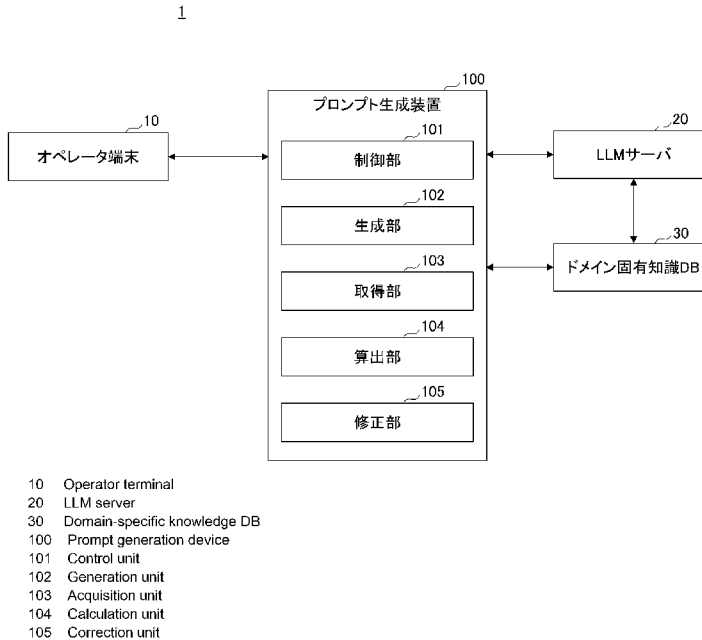
WO 2025/009035 A1

- (51) 国際特許分類:
G06F 40/56 (2020.01) *G06F 40/44* (2020.01)
- (21) 国際出願番号: PCT/JP2023/024678
- (22) 国際出願日: 2023年7月3日(03.07.2023)
- (25) 国際出願の言語: 日本語
- (26) 国際公開の言語: 日本語
- (71) 出願人: 日本電信電話株式会社 (NIPPON TELEGRAPH AND TELEPHONE CORPORATION) [JP/JP]; 〒1008116 東京都千代田区大手町一丁目5番1号 Tokyo (JP).
- (72) 発明者: 高橋 洋介 (TAKAHASHI, Yosuke); 〒1808585 東京都武蔵野市緑町3丁目9-1 NTT知的財産センタ内 Tokyo (JP).
- (74) 代理人: 伊東 忠重, 外 (ITO, Tadashige et al.); 〒1000005 東京都千代田区丸の内二丁目1番1号 丸の内 M Y P L A Z A (明治安田生命ビル) 16階 Tokyo (JP).
- (81) 指定国(表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU,

(54) Title: PROMPT GENERATION DEVICE AND PROMPT GENERATION METHOD

(54) 発明の名称: プロンプト生成装置、及びプロンプト生成方法

[図1]



(57) Abstract: To enable a prompt generation device for generating a prompt to be input to a large language model to easily generate a prompt that produces a more consistent response from the large language model, a prompt generation device according to the present invention includes: a creation unit that creates a prompt to be input to a large language model on the basis of a question or problem for a predetermined event; an acquisition unit that acquires a plurality of responses by inputting the prompt to the large language model a plurality of times; a calculation unit that calculates an

[続葉有]

LY, MA, MD, MG, MK, MN, MU, MW, MX, MY,
MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL,
PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK,
SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA,
UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

- (84) 指定国(表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類：

一 国際調査報告 (条約第21条(3))

evaluation index indicating consistency of the plurality of responses; and a correction unit that corrects the prompt so that a higher evaluation index can be obtained if the evaluation index is less than a predetermined threshold.

(57) 要約：大規模言語モデルに入力するプロンプトを生成するプロンプト生成装置において、大規模言語モデルによる回答の一貫性がより高いプロンプトを容易に生成できるようにするために、プロンプト生成装置は、所定の事象に対する質問又は問題に基づいて、大規模言語モデルに入力するプロンプトを作成する作成部と、前記プロンプトを前記大規模言語モデルに複数回入力して複数の回答を取得する取得部と、前記複数の回答の一貫性を示す評価指標を算出する算出部と、前記評価指標が所定の閾値未満である場合、より高い前記評価指標が得られるように前記プロンプトを修正する修正部と、を有する。

明 細 書

発明の名称： プロンプト生成装置、及びプロンプト生成方法

技術分野

[0001] 本発明は、プロンプト生成装置、及びプロンプト生成方法に関する。

背景技術

[0002] GPT-4 (Generative Pre-trained Transformer-4) に代表される大規模言語モデル (LLM: Large Language Model) が知られている (例えば、非特許文献1 参照)。

[0003] また、質問やタスクを効果的にフォーマットして、大規模言語モデルからより望ましい結果を得るための手法であるプロンプトエンジニアリングが知られている (例えば、非特許文献2 参照)。

先行技術文献

非特許文献

[0004] 非特許文献1: OpenAI, "GPT-4 Technical Report", arXiv preprint arXiv:2303.08774 (2023), <<https://arxiv.org/pdf/2303.08774.pdf>>, インターネット, [令和5年6月23日検索]

非特許文献2: Yang, Xiaohan, et al., "What gpt knows about who is who", arXiv preprint arXiv:2205.07407 (2022), <<https://arxiv.org/pdf/2205.07407.pdf>>, インターネット, [令和5年6月23日検索]

発明の概要

発明が解決しようとする課題

[0005] 大規模言語モデルをネットワーク運用に対処するレコメンドシステムとして適用し、運用者が入力したネットワーク運用の対処方法の質問に対し、その対処方法を回答させることにより、ネットワーク運用を効率化し、運用者の負荷を低減したいという要求がある。

[0006] しかし、従来の大規模言語モデルでは、同一の質問をした場合においても回答内容に一貫性がなく、同一の事象に対して統一した運用の対処方法を提

供することには困難を伴っていた。

[0007] 本発明の実施形態は、上記の課題に鑑みてなされたものであって、大規模言語モデルに入力するプロンプトを生成するプロンプト生成装置において、大規模言語モデルによる回答の一貫性がより高いプロンプトを容易に生成できるようにする。

課題を解決するための手段

[0008] 上記の課題を解決するため、本発明の実施形態に係るプロンプト生成装置は、所定の事象に対する質問又は問題に基づいて、大規模言語モデルに入力するプロンプトを作成する作成部と、前記プロンプトを前記大規模言語モデルに複数回入力して複数の回答を取得する取得部と、前記複数の回答の一貫性を示す評価指標を算出する算出部と、前記評価指標が所定の閾値未満である場合、より高い前記評価指標が得られるように前記プロンプトを修正する修正部と、を有する。

発明の効果

[0009] 本発明の実施形態によれば、大規模言語モデルに入力するプロンプトを生成するプロンプト生成装置において、大規模言語モデルによる回答の一貫性がより高いプロンプトを容易に生成できるようになる。

図面の簡単な説明

[0010] [図1]本実施形態に係る情報処理システムの構成例を示す図である。

[図2]本実施形態に係るプロンプト生成装置の処理の例を示すフローチャートである。

[図3]本実施形態に係る初期プロンプトの例を示す図である。

[図4]本実施形態に係るプロンプト修正のためのプロンプトの例を示す図である。

[図5]本実施形態に係る修正後のプロンプトの例を示す図である。

[図6]本実施形態に係るコンピュータのハードウェア構成の例を示す図である。

発明を実施するための形態

[0011] 以下、図面を参照して本発明の実施の形態（本実施形態）を説明する。以下で説明する実施形態は一例に過ぎず、本発明が適用される実施形態は、以下の実施形態に限られない。

[0012] <情報処理システムの構成>

図1は、本実施形態に係る情報処理システムの構成例を示す図である。情報処理システム1は、例えば、オペレータ端末10、LLM（Large Language Model）サーバ20、ドメイン固有知識DB（Database）30、及びプロンプト生成装置100等を含む。プロンプト生成装置100は、オペレータ端末10、LLMサーバ20に指示サーバ20、及びドメイン固有知識DB30等と通信可能に接続されている。

[0013] 情報処理システム1は、例えば、オペレータ端末10から入力された、所定の事象に対する質問または問題に対して、大規模言語モデル（以下、LLMと呼ぶ）を用いて適切な回答を提供するシステムである。

[0014] ここで、LLMは、大量のテキストデータを使ってトレーニングされた自然言語処理のモデルであり、利用者の質問に対して適切な回答を返すことができる。このLLMをネットワーク（通信ネットワーク）の運用に応用することで、運用者の負荷を低減することが期待できる。例えば、故障対応等のネットワーク運用の対処が必要な場面において、運用者がLLMに対処方法を質問として入力し、LLMは対処方法を回答する。このように活用することで、LLMをネットワークの運用に対処するレコメンドシステムとして使うことが可能となる。

[0015] ここでは、情報処理システム1が、ネットワーク運用の対処方法を回答するレコメンドシステムであるものとして、以下の説明を行う。ただし、これに限られず、情報処理システム1は、ネットワーク運用以外の様々な事象に対する質問又は問題に対する対処方法を回答するシステムであってもよい。

[0016] オペレータ端末10は、運用者（オペレータ）が利用する、例えば、PC（Personal Computer）、タブレット端末、又はスマートフォン等のコンピュ

ータの構成を備えた情報端末である。本実施形態では、運用者が、オペレータ端末10に、ネットワーク運用に関する質問又は問題を入力し、オペレータ端末10は、入力された質問又は問題をプロンプト生成装置100に送信する。

[0017] プロンプト生成装置（プロンプト生成サーバ）100は、例えば、コンピュータの構成を備えた情報処理装置、又は複数のコンピュータを含むシステムである。プロンプト生成装置100は、オペレータ端末10から送信された質問又は問題を受信し、プロンプト生成アルゴリズムを用いて最適なプロンプトを生成する。ここで、プロンプトとは、LLMに対して出力する指示情報であり、例えば、指示文、条件、又は出力形式の指定等の情報を含む。なお、プロンプトは、AI（Artificial Intelligence）プロンプトとも呼ばれる。

[0018] プロンプト生成装置100は、LLMサーバ20と接続されており、LLMサーバ20からの回答内容を活用しながらプロンプト修正を繰り返すことで、最適なプロンプトを生成する。また、プロンプト生成装置100は、ドメイン固有知識DB30とも接続されており、ドメインごとに適切なプロンプトの修正指針に従ってプロンプト修正のためのプロンプトを作成する。

[0019] LLMサーバ20は、プロンプト生成装置100から送信されたプロンプトを受け取り、例えば、GPT-4等の大規模言語モデルを用いて運用者の質問又は問題に対する回答を生成する。また、LLMサーバ20は、ドメイン知識DB30と接続されており、回答を生成する際にはドメイン知識（例えば、ネットワークの構成情報等）を踏まえた適切な回答を作成する。なお、以下の説明において、LLMサーバ20を単に「LLM」と呼ぶ場合がある。

[0020] ドメイン固有知識DB30は、本実施形態では、ネットワークの運用に関連するドメイン固有の知識が格納されているデータベースである。ドメイン固有知識DB30は、LLMサーバ20に対してドメイン知識を提供し、回答の精度を向上させる。また、ドメイン固有知識DB30には、プロンプト

の修正指針も保管されており、プロンプト生成装置１００は、この修正指針に従ってプロンプト修正のためのプロンプトを作成する。

[0021] （プロンプト生成装置の機能構成）

プロンプト生成装置１００は、プロンプト生成装置１００が備えるコンピュータが所定のプログラムを実行することにより、例えば、図１に示すような機能構成を実現している。図１の例では、プロンプト生成装置１００は、制御部１０１、生成部１０２、取得部１０３、算出部１０４、及び修正部１０５等を有する。なお、上記の各機能構成のうち、少なくとも一部は、ハードウェアによって実現されるものであってもよい。

[0022] 制御部１０１は、オペレータ端末１０からの質問又は問題を受け付けし、受け付けた質問又はに基づいて生成したプロンプトをＬＬＭサーバ２０に入力し、ＬＬＭサーバ２０から得た回答をオペレータ端末１０に提供する。

[0023] なお、ＬＬＭサーバ２０は、本実施形態に係る大規模言語モデル（ＬＬＭ）の一例である。本実施形態に係る大規模言語モデルは、例えば、外部のクラウドサービス等が提供する大規模言語モデル等であってもよい。

[0024] 生成部１０２は、所定の事象に対する質問又は問題に基づいて、大規模言語モデルに入力するプロンプトを生成する生成処理を実行する。例えば、生成部１０２は、制御部１０１がオペレータ端末１０から受け付けた、ネットワークの運用に関する質問又は問題に基づいて、ＬＬＭサーバ２０に入力する初期プロンプトを生成する。

[0025] 取得部１０３は、プロンプトを大規模言語モデルに複数回入力して複数の回答を取得する取得処理を実行する。例えば、取得部１０３は、生成部１０２が生成したプロンプト、又は修正部１０５が修正した同一のプロンプトを、ＬＬＭサーバ２０に複数回入力して、複数の回答を取得する。

[0026] 算出部１０４は、複数の回答の一貫性を示す評価指標を算出する算出処理を実行する。例えば、算出部１０４は、取得部１０３が取得した複数の回答の一貫性を示す一貫性スコア、又は類似度等を算出する。なお、一貫性スコア、又は類似度は、複数の回答の一貫性を示す評価指標の一例である。

[0027] 修正部 105 は、複数の回答の一貫性を示す評価指標が所定の閾値未満である場合、より高い評価指標が得られるようにプロンプトを修正する修正処理を実行する。例えば、修正部 105 は、算出部 104 が算出した一貫性スコア（又は類似度）が、予め定められた閾値に満たない場合、LLMサーバ 20 等の大規模言語モデルを用いて、より高い一貫性スコア（又は類似度）が得られるように、プロンプトを修正する。

[0028] 好ましくは、修正部 105 は、例えば、ドメイン固有知識 DB 30 に含まれる、ネットワーク運用に関するプロンプトの修正指針を用いて、複数の回答の一貫性を示す評価指標が所定の閾値以上になるまで、プロンプトを繰り返し修正する。

[0029] （従来技術の課題について）

従来の LLM は、同一の質問をした場合においても、回答内容に一貫性がなく、例えば、ネットワークの運用にそのまま適用することは困難である。これは、ネットワーク運用には複数の運用者が関わるため、LLM が一貫性のない回答を返してしまうと、同一の事象に対して各運用者が統一した対処を行うことが難しいためである。

[0030] なお、質問の書き方によっては、一貫性の高い（ぶれの少ない）回答が期待できることが知られているが、運用者が都度適切な質問を作成することは困難である。そこで、自動的に最適なプロンプトを生成する技術が求められている。

[0031] 従来の LLM では、回答の一貫性を向上させるために、主にパラメータ調整とプロンプトエンジニアリングの 2 つのアプローチが取られている。

[0032] パラメータ調整では、LLM の挙動を制御するために、例えば、temperature、又は top_p 等のパラメータを調整することにより、生成されるテキストの多様性、又は一貫性を調整することが可能である（例えば、非特許文献 1 参照）。しかし、この方法は、適切なパラメータチューニングが難しく、誤ったパラメータ設定が回答の精度や多様性を極端に損なう可能性がある。また、この方法は、一部のパラメータが外部化されておらず、ユーザ側で変更が

できない場合もあり、その制限により適切な調整ができないこともある。

[0033] プロンプトエンジニアリングは、質問又タスクを効果的にフォーマットして、LLMからより望ましい結果を得るための手法である（例えば、非特許文献2参照）。プロンプトエンジニアリングでは、例えば、明確な指示、質問の再フォーマット、文脈の追加、ステップバイステップ等の手法が一般的に用いられる。適切にプロンプトエンジニアリングを行うことにより、回答の一貫性、及び精度等を向上させることができる。しかし、運用者が、質問を作成するたびに適切にプロンプトエンジニアリングを手動で実施することは困難である。

[0034] このように、従来のLLMでは、同一の質問をした場合においても回答内容に一貫性がなく、同一の事象に対して統一した運用の対処方法を提供することには困難を伴っていた。

[0035] そこで、本実施形態に係るプロンプト生成装置100は、図1に示すような機能構成により、例えば、ネットワーク運用支援のための自動プロンプト最適化技術を提供する。この技術により、運用者は、LLMから一貫性のある回答を容易に得られるようになり、ネットワーク運用における問題の解決が迅速かつ効率的に行われることが期待できる。また、この技術により、運用者の負担が軽減され、ネットワーク管理の効率化が図られることが期待できる。

[0036] このように、本実施形態によれば、大規模言語モデルに入力するプロンプトを生成するプロンプト生成装置において、大規模言語モデルによる回答の一貫性がより高いプロンプトを容易に生成できるようになる。

[0037] <処理の流れ>

続いて、本実施形態に係るプロンプト生成方法の所定の流れについて説明する。図2は、本実施形態に係るプロンプト生成装置の処理の例を示すフローチャートである。この処理は、例えば、図1で説明した機能構成を有するプロンプト生成装置100が実行する処理の例を示している。

[0038] ステップS201において、生成部102は、所定の事象に対する質問又

は問題に基づいて、大規模言語モデルに入力する初期プロンプトを生成する生成処理を実行する。例えば、生成部102は、制御部101が、オペレータ端末10から入力されたネットワーク運用に関する質問又は問題等に基づいて、LLMサーバ20に入力する初期プロンプトを生成する。

[0039] 初期プロンプトは、対象となる大規模言語モデル（例えば、GPT-4）に対して適切な回答を得ることができるように、質問又は問題の内容を正確かつ簡潔に表現するものである。なお、初期プロンプトは、ユーザからの入力、既存のテンプレート、又は自動生成された質問を元に作成される場合がある。

[0040] 図3は、本実施形態に係る初期プロンプトの例を示す図である。図3の例では、初期プロンプト300は、「イベント情報」と「指示」を含む形で構成されている。この初期プロンプト300は、対象となるLLMに対して、ネットワーク機器の異常トラフィックに関連する情報を適切に抽出し、回答を生成するよう促すものである。

[0041] なお、初期プロンプトの作成にあたっては、対象となる質問又は問題の難易度、背景知識等を考慮し、適切な質問形式、又は表現を選択することが重要である。これにより、対象となるLLMが、質問又は問題を理解し易くなり、回答の一貫性、又は精度等が向上することが期待される。

[0042] ステップS201における初期プロンプトの作成は、本実施形態におけるプロンプトの最適化プロセスの出発点となる。

[0043] ステップS202において、取得部103は、プロンプトをLLMに複数回入力して複数の回答を取得する取得処理を実行する。例えば、取得部103は、生成部102が生成した初期プロンプト、又は修正部105が修正した修正後のプロンプトをLLMサーバ20に複数回入力し、プロンプトに対する複数の回答を取得する。ここで、回答を複数回取得する理由は、複数回の回答の一貫性を評価し、プロンプトの改善を行うためである。

[0044] ステップS203において、算出部104は、取得部103が取得した複数の回答の一貫性を示す評価指標を算出する。例えば、算出部104は、公

知のBERTスコアを用いて回答間の類似度を計算し、質問の評価指標として用いる。BERTスコアは、BERT(Bidirectional Encoder Representations from Transformers)モデルと呼ばれる自然言語処理モデルを利用した評価指標であり、文の類似度を0から1の範囲で評価する。BERTスコアが高いほど、回答の一貫性が高いことを示す。(参考文献: Zhang, Tianyi, et al, "Bertscore: Evaluating text generation with bert", arXiv preprint arXiv: 1904.09675 (2019), <<https://arxiv.org/pdf/1904.09675.pdf>>)

[0045] 算出部104は、回答が3つ以上の場合、全ての回答ペアのBERTスコア(0から1の実数、数値が大きいほど類似度が高い)の平均値を、その回答の一貫性スコアとする。

[0046] 例えば、算出部104は、次の手順でBERTスコアを計算する。

(1) 算出部104は、生成された各回答を、BERTモデルに入力し、文の埋め込みベクトルを取得する。

(2) 算出部104は、全ての回答ペアについて、それぞれの埋め込みベクトル間のコサイン類似度を算出する。

(3) 算出部104は、算出されたコサイン類似度を0から1の範囲にスケールリングし、各回答ペアのBERTスコアを得る。

(4) 算出部104は、全ての回答ペアのBERTスコアの平均値を、複数の回答の一貫性スコアとする。

[0047] ステップS204において、修正部105は、算出部104が算出した複数の回答の一貫性を示す評価指標が、所定の閾値以上であるか否かを判断する。ここで、所定の閾値は、複数の回答の一貫性が十分であるか否かを判断するための値が、予め設定されているものとする。

[0048] また、この閾値は、実際のタスク、又は応用シーン等に応じて調整が可能である。例えば、この閾値を低く設定すれば、一貫性の低い回答を許容することになり、少ない繰り返し回数で回答が得やすくなる。一方、この閾値を高く設定すれば、その逆となり、最終的な回答を得るまでの繰り返し回数が多くなる。

- [0049] 複数の回答の一貫性を示す評価指標が、所定の閾値未満である場合、修正部105は、現在のプロンプトに対する複数の回答の一貫性が低いと判断し、処理をステップS205に移行させる。一方、複数の回答の一貫性を示す評価指標が、所定の閾値以上である場合、修正部105は、現在のプロンプトに対する複数の回答の一貫性が十分に高いと判断し、処理をステップS206に移行させる。
- [0050] ステップS205に移行すると、修正部105は、より高い評価指標が得られるように、プロンプトを修正し、処理をステップS202に戻す。好ましくは、修正部105は、LLMを用いて、プロンプトを修正する。
- [0051] 例えば、修正部105は、プロンプトの内容と修正指針をLLMサーバ20に入力し、より一貫性の高い回答を得られるようにプロンプトを修正させる。例えば、修正部105は、図4に示すような、プロンプト修正のためのプロンプトを作成し、LLMサーバ20に入力する。
- [0052] 図4は、本実施形態に係るプロンプト修正のためのプロンプトの例を示す図である。図4の例では、プロンプト修正のためのプロンプト400は、修正指針401を含む。修正指針401は、プロンプトエンジニアリングの考え方を応用して、タスクや利用シーンに適したものを予め作成しておく。図4の例では、修正指針401は、「質問の目的および期待する回答を明確にする」、「質問の内容を具体的にする」、及び「回答に含める情報やフォーマットを指定する」等の指針を含む。修正部105は、これらの指針に従って、LLMにプロンプトを改善させることにより、より一貫性の高い回答が得られるように、プロンプトを修正することができる。
- [0053] 変形例として、修正部105は、プロンプト修正のためのプロンプトを複数用意しておき、一貫性を示す評価指標の値に応じて、修正指針を変更してもよい。例えば、一貫性を示す評価指標の値が小さい場合、修正部105は、プロンプトの大幅な修正を指示するプロンプトを用いてもよい。また、修正部105は、一貫性を示す評価指標の改善度合いが小さい場合、これまで使用していたプロンプト修正のためのプロンプトとは、指示の内容が大きく

異なる別のプロンプト修正のためのプロンプトを試してもよい。また、修正部105は、LLMサーバ20とは別のLLMを用いて、プロンプトを修正してもよい。

[0054] 修正部105は、ステップS204の判断結果に基づいて、プロンプトの修正プロセスを繰り返すことが可能であり、適切な回答が得られるまで続けることができる。ただし、無限回の修正が行われることを防ぐため、修正部105は、修正回数に上限を設定し、その上限に達した場合にはプロンプトの修正プロセスを終了することが望ましい。また、修正部105は、複数回の修正を実施しても一貫性を示す評価指標が高くない場合、初期プロンプト、又は修正指針を見直して、処理を再実行することがのぞましい。これにより、より効率的なプロンプトの修正プロセスが実現される。

[0055] 図5は、本実施形態に係る修正後のプロンプトの例を示す図である。図5の例では、修正後のプロンプト500は、図3で説明した初期プロンプト300と同様に、「イベント情報」と「指示」を含む形で構成されている。また、修正後のプロンプト500に含まれる「指示」は、修正指針401に従って、より一貫性の高い回答が得られるように改善されている。

[0056] ステップS206に移行すると、制御部101は、ステップS201～S205の処理で生成したプロンプトを用いて取得した回答を提供する。例えば、制御部101は、オペレータ端末10から受け付けた質問又は問題に基づいて、ステップS201～S205の処理で生成したプロンプトを、LLMサーバ20に入力する。また、制御部101は、LLMサーバ20から取得した回答を、オペレータ端末10に提供する。

[0057] 図2の処理により、大規模言語モデルに入力するプロンプトを生成するプロンプト生成装置100において、大規模言語モデルによる回答の一貫性がより高いプロンプトを容易に生成できるようになる。

[0058] <ハードウェア構成例>

プロンプト生成装置100は、例えば、図6に示すようなコンピュータ600のハードウェア構成を有している。或いは、プロンプト生成装置100

は、複数のコンピュータ600によって構成される。

[0059] 図6は、本実施形態に係るコンピュータのハードウェア構成の例を示す図である。プロンプト生成装置100は、例えば、図5に示すようなコンピュータ600のハードウェア構成を有している。図6の例では、コンピュータ600は、プロセッサ601、メモリ602、ストレージデバイス603、通信装置604、入力装置605、出力装置606、及びバスB等を有する。

[0060] プロセッサ601は、例えば、所定のプログラムを実行することにより、様々な機能を実現するCPU (Central Processing Unit) 等の演算装置である。メモリ602は、コンピュータ600が読み取り可能な記憶媒体であり、例えば、RAM (Random Access Memory)、ROM (Read Only Memory) 等を含む。ストレージデバイス603は、コンピュータ読み取り可能な記憶媒体であり、例えば、HDD (Hard Disk Drive)、SSD (Solid State Drive)、各種の光ディスク、及び光磁気ディスク等を含み得る。

[0061] 通信装置604は、無線、又は有線のネットワークを介して他の装置と通信を行うための1つ以上のハードウェア（通信デバイス）を含む。入力装置605は、外部からの入力を受け付ける入力デバイス（例えば、キーボード、マウス、マイクロフォン、スイッチ、ボタン、センサ等）である。出力装置606は、外部への出力を実施する出力デバイス（例えば、ディスプレイ、スピーカ、LEDランプ等）である。なお、入力装置605と出力装置606とは、一体となった構成（例えば、タッチパネルディスプレイ等の入出力装置）であってもよい。

[0062] バスBは、上記の各構成要素に共通に接続され、例えば、アドレス信号、データ信号、及び各種の制御信号等を伝送する。なお、プロセッサ601は、CPUに限られず、例えば、DSP (Digital Signal Processor)、PLD (Programmable Logic Device)、又はFPGA (Field Programmable Gate Array) 等であってもよい。

[0063] （補足）

本実施形態におけるプロンプト生成装置１００は、専用装置による実現に限らず、汎用コンピュータで実現するようにしてもよい。その場合、この機能を実現するためのプログラムをコンピュータ読み取り可能な記録媒体に記録して、この記録媒体に記録されたプログラムをコンピュータシステムに読み込ませ、実行することによって実現してもよい。なお、ここでいう「コンピュータシステム」とは、ＯＳや周辺機器等のハードウェアを含むものとする。

[0064] また、「コンピュータ読み取り可能な記録媒体」とは、フレキシブルディスク、光磁気ディスク、ＲＯＭ、ＣＤ－ＲＯＭ等の可搬媒体、コンピュータシステムに内蔵されるストレージデバイス６０３等の様々な記憶装置を含む。さらに「コンピュータ読み取り可能な記録媒体」とは、インターネット等のネットワークや電話回線等の通信回線を介してプログラムを送信する場合の通信線のように、短時間の間、動的にプログラムを保持するもの、その場合のサーバやクライアントとなるコンピュータシステム内部の揮発性メモリのように、一定時間プログラムを保持しているものも含んでもよい。

[0065] また上記プログラムは、前述した機能の一部を実現するためのものであっても良く、さらに前述した機能をコンピュータシステムにすでに記録されているプログラムとの組み合わせで実現できるものであっても良く、ＰＬＤ（Programmable Logic Device）やＦＰＧＡ（Field Programmable Gate Array）等のハードウェアを用いて実現されるものであってもよい。

[0066] ＜実施形態の効果＞

本実施形態によって生じる効果は、以下の通りである。

（１）ＬＬＭ（大規模言語モデル）による回答内容の一貫性の向上:本実施形態によるプロンプトの改善を通じて、ＬＬＭが生成する回答の一貫性が向上し、ユーザにとって混乱を招くことが減少する。

（２）プロンプトの明確化によるＬＬＭの回答精度の向上:プロンプトの明確化により、ＬＬＭが適切な回答を生成し易くなり、回答精度が向上する。これにより、ユーザが必要な情報にアクセスする際の効率が向上する。

(3) 効率的なプロンプト修正による適切な回答取得の促進:本実施形態により、プロンプトを効率的に修正できるようになり、適切な回答を取得する確率が高まる。これにより、ユーザは迅速に問題を解決できるようになる。

(4) 運用者の負担軽減:本実施形態の自動プロンプト最適化技術は、運用者が都度適切な質問を作成する困難さを軽減し、運用者の負担を軽減する。

(5) ネットワーク管理の効率化:本実施形態により、運用者が簡単に一貫性のある回答を得ることができるようになり、ネットワーク運用における問題の解決が迅速かつ効率的に行われることが期待される。これにより、ネットワーク管理の効率化が図られる。

(6) LLMの応用範囲の拡大:本実施形態による情報処理システム1の品質向上は、LLMの応用範囲を拡大させる可能性がある。これにより、新たな応用分野でのLLMの利用が促進されることが期待される。

[0067] 以上の効果により、本実施形態によれば、情報処理システム1における回答品質の向上、又はユーザの情報ニーズに対応する能力を強化することができる。また、本実施形態によれば、ネットワーク運用における問題の解決が迅速かつ効率的に行われ、運用者の負担が軽減される。これにより、組織全体の生産性が向上し、ネットワーク管理がより効率的に実施されることが期待される。さらに、本実施形態が他の業界や応用分野に応用されることで、さらなる効果が発揮される可能性がある。

[0068] <本実施形態のポイント>

本実施形態の技術的なポイントは以下の通りである。

(1) 自動プロンプト生成アルゴリズム:本実施形態では、収集、分析されたデータを基に、自動的に適切なプロンプトを生成するアルゴリズムを開発している。このアルゴリズムは、質問のタイプ、又は回答の目的等に応じて最適なプロンプトを提案し、運用者の手間を軽減する。

(2) プロンプト評価機能:本実施形態には、生成されたプロンプトの効果を評価する機能が含まれている。この機能により、プロンプトの質が継続的に向上し、運用者がより効果的な質問を行えるようになる。

(3) 反復的最適化プロセス: 本実施形態は、生成されたプロンプトとその評価結果を基に、プロンプト生成アルゴリズムを反復的に実行して最適化を行う。これにより、常に最適なプロンプトが提供され、ネットワーク運用支援が効率的に行われる。

(4) ドメイン固有の知識の導入: 本実施形態は、ネットワーク運用に関連するドメイン固有の知識を取り込むことができる。これにより、運用者が専門的な質問を行った際にも適切な回答が得られ、より効果的な運用支援が実現される。

[0069] <実施形態のまとめ>

本明細書には、少なくとも下記の各項のプロンプト生成装置、及びプロンプト生成方法が開示されている。

(第1項)

大規模言語モデルに入力するプロンプトを生成する生成部と、
前記プロンプトを前記大規模言語モデルに複数回入力して複数の回答を取得する取得部と、
前記複数の回答の一貫性を示す評価指標を算出する算出部と、
前記評価指標が所定の閾値未満である場合、より高い前記評価指標が得られるように前記プロンプトを修正する修正部と、
を有する、プロンプト生成装置。

(第2項)

前記修正部は、前記大規模言語モデルを用いて前記プロンプトを修正する、第1項に記載のプロンプト生成装置。

(第3項)

前記プロンプトは、ネットワーク運用における事象に対する質問又は問題を含み、
前記修正部は、前記ネットワーク運用に関する前記プロンプトの修正指針を用いて、前記評価指標が前記所定の閾値以上となるまで、前記プロンプトを繰り返し修正する、

第1項又は第2項に記載のプロンプト生成装置。

(第4項)

大規模言語モデルに入力するプロンプトを生成する処理と、

前記プロンプトを前記大規模言語モデルに複数回入力して複数の回答を取得する処理と、

前記複数の回答の一貫性を示す評価指標を算出する処理と、

前記評価指標が所定の閾値未満である場合、より高い前記評価指標が得られるように前記プロンプトを修正する処理と、

をコンピュータが実行する、プロンプト生成方法。

[0070] (第5項)

大規模言語モデルに入力するプロンプトを生成する処理と、

前記プロンプトを前記大規模言語モデルに複数回入力して複数の回答を取得する処理と、

前記複数の回答の一貫性を示す評価指標を算出する処理と、

前記評価指標が所定の閾値未満である場合、より高い前記評価指標が得られるように前記プロンプトを修正する処理と、

をコンピュータに実行させる、プログラムを記憶した記憶媒体。

[0071] 以上、本実施形態について説明したが、本発明はかかる特定の実施形態に限定されるものではなく、特許請求の範囲に記載された本発明の要旨の範囲内において、種々の変形・変更が可能である。

符号の説明

- [0072] 1 情報処理システム
- 10 オペレータ端末
 - 20 LLMサーバ
 - 30 ドメイン固有知識DB
 - 101 制御部
 - 102 生成部
 - 103 取得部

1 0 4 算出部

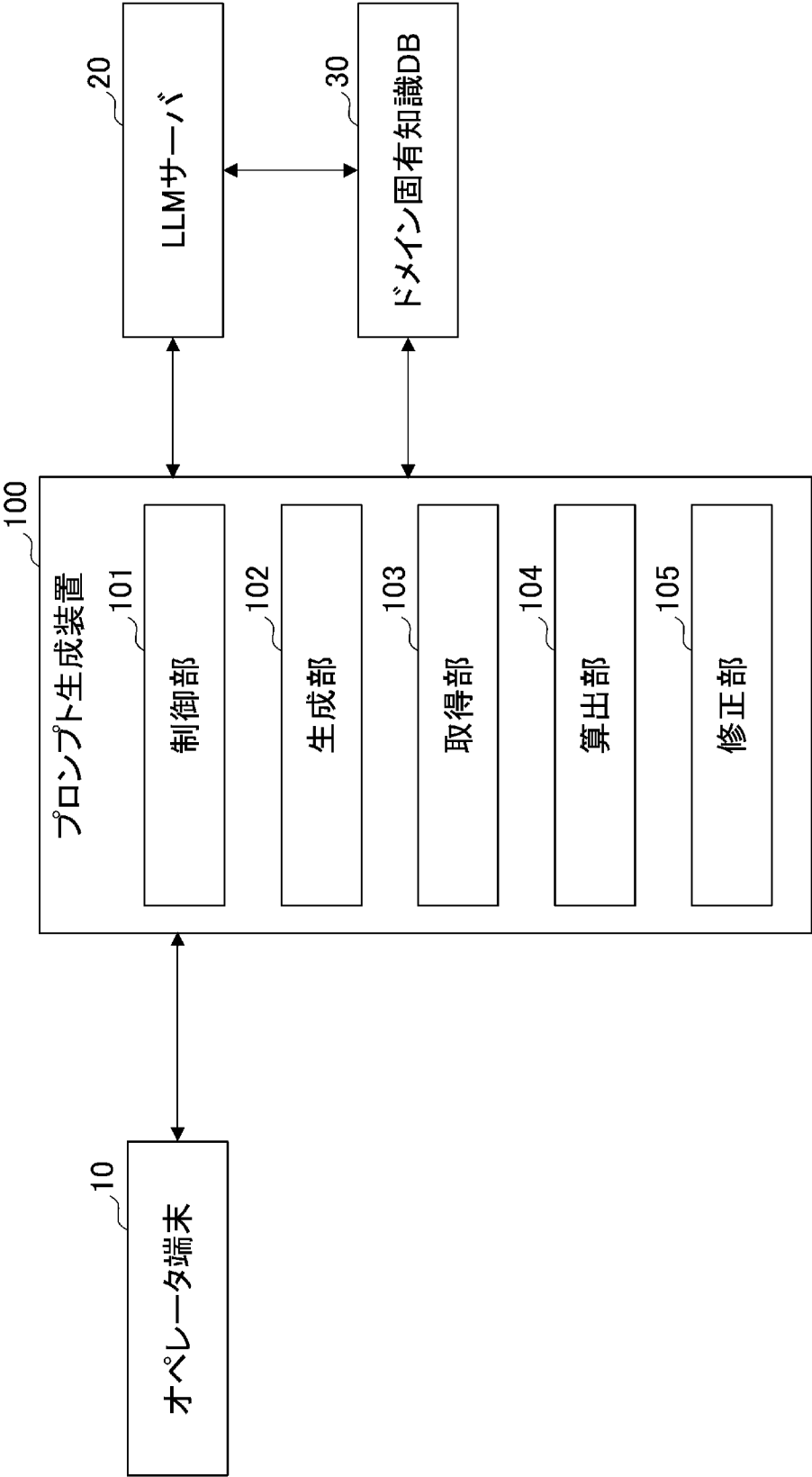
1 0 5 修正部

請求の範囲

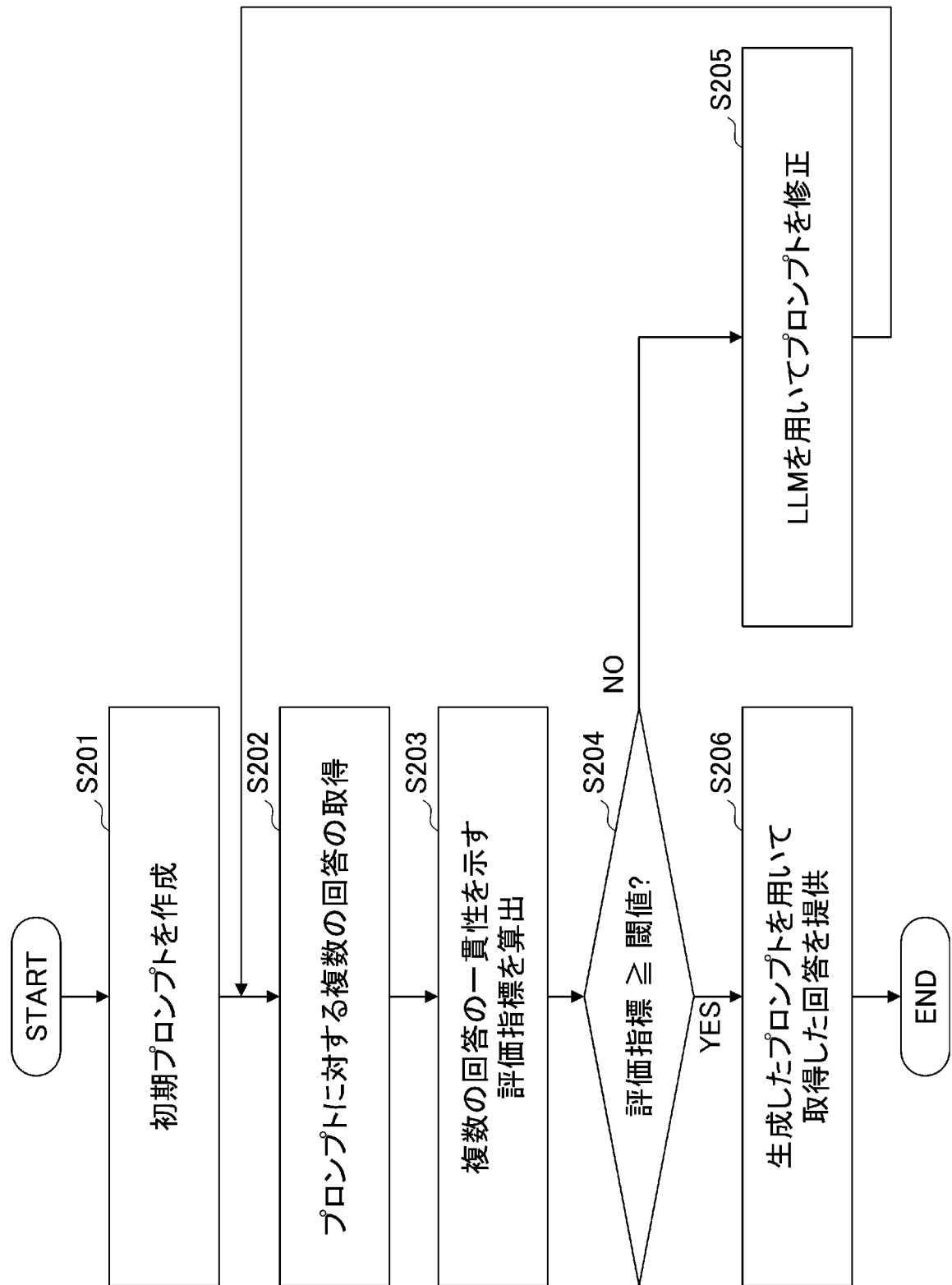
- [請求項1] 大規模言語モデルに入力するプロンプトを生成する生成部と、
前記プロンプトを前記大規模言語モデルに複数回入力して複数の回答を取得する取得部と、
前記複数の回答の一貫性を示す評価指標を算出する算出部と、
前記評価指標が所定の閾値未満である場合、より高い前記評価指標が得られるように前記プロンプトを修正する修正部と、
を有する、プロンプト生成装置。
- [請求項2] 前記修正部は、前記大規模言語モデルを用いて前記プロンプトを修正する、請求項1に記載のプロンプト生成装置。
- [請求項3] 前記プロンプトは、ネットワーク運用における事象に対する質問又は問題を含み、
前記修正部は、前記ネットワーク運用に関する前記プロンプトの修正指針を用いて、前記評価指標が前記所定の閾値以上となるまで、前記プロンプトを繰り返し修正する、
請求項1又は2に記載のプロンプト生成装置。
- [請求項4] 大規模言語モデルに入力するプロンプトを生成する処理と、
前記プロンプトを前記大規模言語モデルに複数回入力して複数の回答を取得する処理と、
前記複数の回答の一貫性を示す評価指標を算出する処理と、
前記評価指標が所定の閾値未満である場合、より高い前記評価指標が得られるように前記プロンプトを修正する処理と、
をコンピュータが実行する、プロンプト生成方法。

[図1]

1



[図2]



[図3]

300

[イベント情報]

日時: 2023-04-24 10:20

イベント: 異常トラフィック検出

送信元IP: 192.168.1.1

宛先IP: 192.168.1.2

プロトコル: UDP

ポート: 123

トラフィック量: 2.5 Gbps

[指示]

上記の異常トラフィックを検出した際の対処法を提案してください。特に、ネットワーク設定の変更やファイアウォールルールの追加に重点を置いて提案してください。

[図4]

400

下記の質問をLLMサーバに3回送信した場合、それぞれ異なる回答が返されました。

回答のブレが小さくなるように[指示]を修正してください。

401

なお、回答のブレが小さくなるような質問修正の指針は、「質問の目的や期待する

回答を明確にする」、「質問の内容を具体的ににする」、「回答に含める情報やフォーマットを指定する」です。

[イベント情報]と修正後の[指示]だけを表示してください。

[イベント情報]

日時: 2023-04-24 10:20

イベント: 異常トラフィック検出

送信元IP: 192.168.1.1

宛先IP: 192.168.1.2

プロトコル: UDP

ポート: 123

トラフィック量: 2.5 Gbps

[指示]

上記の異常トラフィックを検出した際の対処法を提案してください。特に、ネットワーク設定の変更やファイアウォールルールの追加に重点を置いて提案してください。

[図5]

500

[イベント情報]

日時: 2023-04-24 10:20

イベント: 異常トラフィック検出

送信元IP: 192.168.1.1

宛先IP: 192.168.1.2

プロトコル: UDP

ポート: 123

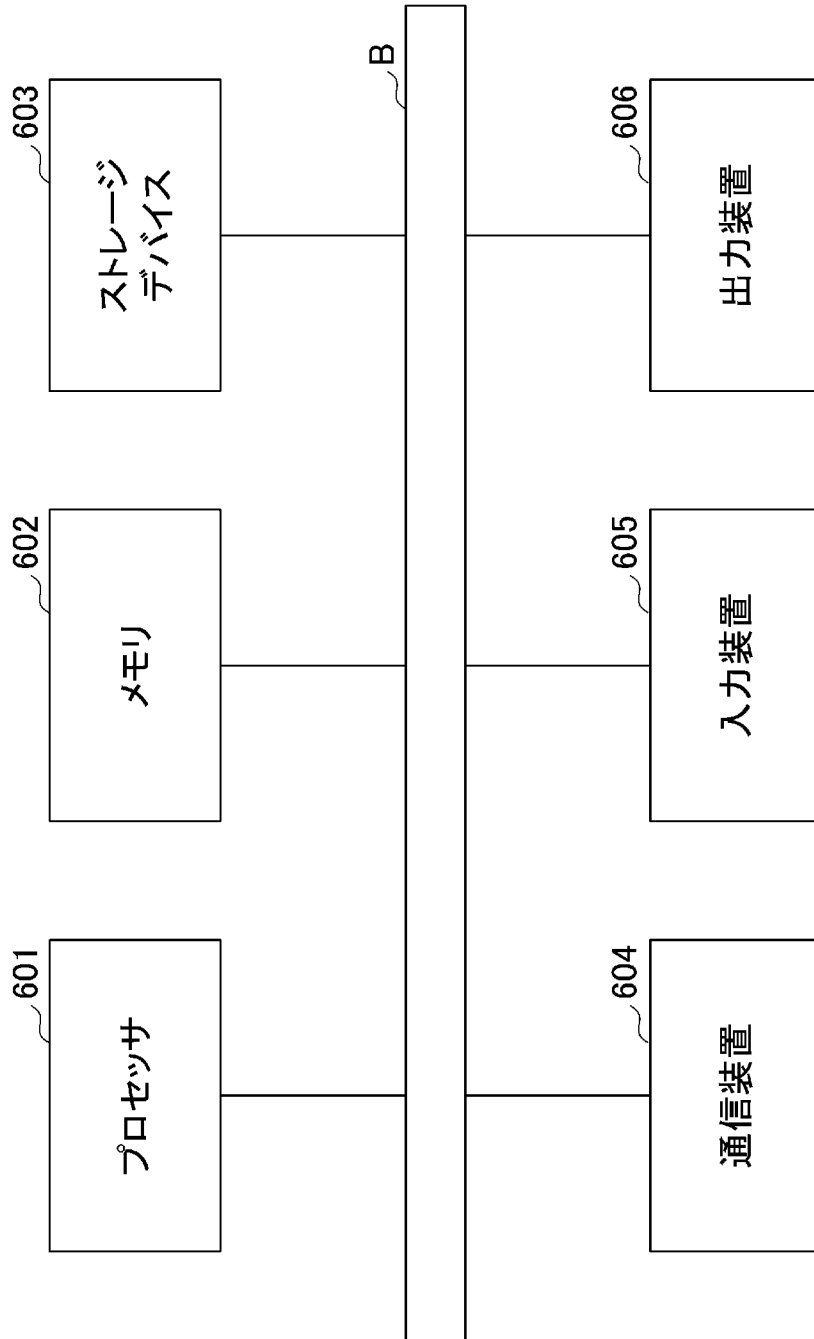
トラフィック量: 2.5 Gbps

[指示]

以下の観点に基づき、上記の異常トラフィックに対する対処法を詳細に提案してください:

- 1) ネットワーク設定の変更に関する具体的な手順,
 - 2) ファイアウォールルールの追加または変更に関する具体的な手順,
 - 3) セキュリティ上の影響を考慮した対策,
 - 4) 現象が再発しないようにするための長期的な対策.
- この4つの観点を明確に区分し、それぞれに対する提案を行ってください。また提案はUDPプロトコルを使用するトラフィックに特化したものであること、そしてポート123を利用することを考慮に入れてください。

[図6]

600

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2023/024678

A. CLASSIFICATION OF SUBJECT MATTER**G06F 40/56**(2020.01)i; **G06F 40/44**(2020.01)i

FI: G06F40/56; G06F40/44

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F40/56; G06F40/44

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Published examined utility model applications of Japan 1922-1996
 Published unexamined utility model applications of Japan 1971-2023
 Registered utility model specifications of Japan 1996-2023
 Published registered utility model applications of Japan 1994-2023

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	MADAAN, A. et al. MemPrompt: Memory-assisted Prompt Editing with User Feedback. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics [online]. 11 December 2022, pp. 2833-2861, [retrieved on 31 July 2023], internet:<URL: https://aclanthology.org/2022.emnlp-main.183.pdf > entire text, all drawings	1-4

☐ Further documents are listed in the continuation of Box C.☐ See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

31 July 2023

Date of mailing of the international search report

08 August 2023

Name and mailing address of the ISA/JP

Japan Patent Office (ISA/JP)
 3-4-3 Kasumigaseki, Chiyoda-ku, Tokyo 100-8915
 Japan

Authorized officer

Telephone No.

A. 発明の属する分野の分類（国際特許分類（IPC）） G06F 40/56(2020.01)i; G06F 40/44(2020.01)i FI: G06F40/56; G06F40/44		
B. 調査を行った分野		
調査を行った最小限資料（国際特許分類（IPC）） G06F40/56; G06F40/44		
最小限資料以外の資料で調査を行った分野に含まれるもの 日本国実用新案公報 1922 - 1996年 日本国公開実用新案公報 1971 - 2023年 日本国実用新案登録公報 1996 - 2023年 日本国登録実用新案公報 1994 - 2023年		
国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）		
C. 関連すると認められる文献		
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号
A	MADAAN, A, ほか, MemPrompt: Memory-assisted Prompt Editing with User Feedback, Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics [online], 2022.12.11, pp. 2833 - 2861, [検索日 2023.07.31], インターネット <URL: https://aclanthology.org/2022.emnlp-main.183.pdf > 全文, 全図	1-4
☐ C欄の続きにも文献が列挙されている。 ☐ パテントファミリーに関する別紙を参照。		
* 引用文献のカテゴリー “A” 特に関連のある文献ではなく、一般的技术水準を示すもの “E” 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの “L” 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す） “O” 口頭による開示、使用、展示等に言及する文献 “P” 国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献	“T” 国際出願日又は優先日後に公表された文献であって出願と抵触するものではなく、発明の原理又は理論の理解のために引用するもの “X” 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの “Y” 特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの “&” 同一パテントファミリー文献	
国際調査を完了した日 31.07.2023	国際調査報告の発送日 08.08.2023	
名称及びあて先 日本国特許庁(ISA/JP) 〒100-8915 日本国 東京都千代田区霞が関三丁目4番3号	権限のある職員（特許庁審査官） 木村 大吾 5N 4807 電話番号 03-3581-1101 内線 3586	