F-032

動的プロプロンプト更新を用いた顧客インタビュートレーニングシステムにおける情報一貫性の検証と応答パターン分析

Verification of Information Consistency and Response Pattern Analysis in Customer Interview Training System Using Dynamic Prompt Updating

川島 壮生† 櫻井 崇貴† 長澤 史記† 白松 俊†
Soki Kawashima Takayoshi Sakurai Fuminori Nagasawa Shun Shiramatsu

1. はじめに

近年，大規模言語モデル（LLM）の急速な発展により，人間の代理や対話相手としてAIを活用する場面が飛躍的に増加している．カスタマーサポート，教育現場でのバーチャル講師，企業研修でのロールプレイング相手，さらには心理カウンセリングの補助ツールに至るまで，LLMは単なる情報提供ツールの枠を超え，人間らしい対話を行う「対話主体」としての役割を担うことが期待されている．こうした応用分野では，AIに求められるのは正確な情報提供だけでなく，文脈に応じた自然な応答，共感的な反応，そして人間らしい振る舞いである．

しかし，従来の静的なプロンプト設計には限界があることが明らかになってきた[1]．静的プロンプトとは，対話開始前にあらかじめ与えられた固定的な指示文や情報であり，その内容に基づいて LLM が応答を生成する．そのような静的プロンプトでは，事前に設定された情報のみに基づく画一的な応答しか生成できず，対話の進行や文脈の変化に柔軟に対応することができない．

そこで我々は先行研究において，対話の進行に応じてプロンプトをリアルタイムで更新する手法を開発した[1]．具体的には，顧客情報を 3 段階の階層構造で管理し，ユーザが質問を通じてシステムから情報を聞き出すたびに，追加情報をプロンプトに段階的に追記するシステムである．例えば，ノンアルコールビールに関するインタビューでは，最初は「健康のためにビールを飲む量を減らしたが満足感が得られなかった」という表面的な情報のみを LLM のプロンプトに与え，ユーザが適切な深堀り質問により LLM から情報を聞き出した場合にのみ「満足感とはストレス発散時の開放感」「そのストレスは主に仕事で感じる」といった，より深層の情報を段階的にプロンプトに追加していく．この仕組みにより，顧客が対話の中で自らの潜在意識を徐々に認識するプロセスを再現することができる．

この動的プロンプト手法の有効性は実験によって実証されている．従来の静的プロンプト手法ではインタビュアーとしてシステムを利用したユーザーからの質問数が平均35%の増加にとどまったのに対し，動的プロンプト手法では平均148%の増加を実現し，1つの情報を聞き出すために要する質問数も最大で約 4 倍となった．これにより，ユーザーがより深い質問スキルを身につけるという訓練目的が効果的に達成された．

しかしながら，動的プロンプトの導入は新たな技術的課題も生み出している．最も深刻な問題の一つが，システムが途中で更新したプロンプトと，LLMがすでに出力した過去の発話との間に生じる矛盾である．動的プロンプト手法では対話の進行に応じて新しい情報がプロンプトに追加されるため，AIが過去に発言した内容と，新たに追加されたプロンプト情報との間で整合性が取れなくなる状況が生じうる．

こうした現象を，本研究では過去の発話内容と外部から与えられる情報源（プロンプト）との間に生じる矛盾という意味で「外的矛盾」と呼ぶ[2]．なお，これとは対照的に，過去の発話の内部で生じる論理的な不整合は「内的矛盾」として区別される．外的矛盾は，人間の対話においては極めて自然に処理される現象である．人間は「あ，さっき言ったことと違うことを考えた」「記憶違いだった」「新しい情報を思い出した」といった形で，過去の発言を自然に修正したり，考えを更新したりすることができる．このような認知的な柔軟性は，人間らしい対話の重要な特徴の一つである．

一方，LLMが同様の特徴を獲得しているかは明らかではない．プロンプトに記載された新しい情報と過去の発話内容が食い違った場合，LLMはどちらを優先すべきか，どのようにして整合性を回復すべきか，明確な戦略を持たない限り，不自然な応答や一貫性の欠如を招く可能性がある．こうした問題は，ユーザーの信頼を損ない，訓練効果を著しく低下させる要因となりうる．

これらの問題は，生成AIの実用的な利用が進む現在において避けることのできない課題である．今後動的プロンプト技術の現場導入が検討される際に，外的矛盾に起因する問題はますます顕在化することが予想される．企業の研修システム，教育機関での学習支援ツール，医療・福祉分野での対話支援システムなど，高い信頼性と人間らしさが求められる重要な応用分野においては，こうした矛盾処理能力の欠如は致命的な問題となりかねない．

さらに，この問題の解決は単なる技術的な改善にとどまらず，LLMが真に「人間らしい対話エージェント」として機能するための根本的な要件に関わっている．人間のように「自分の過去発言を踏まえつつ，新しい情報や状況に応じて考えを更新・修正する能力」は，自然で信頼できる対話を実現するために重要な特性の一つである．この能力の向上は，今後の対話AI技術の発展において重要な要素となると考えられる．

こうした背景を踏まえ，本研究では，外的矛盾（過去発話とプロンプトが矛盾している状況）に直面した際のLLMの応答特性について，その傾向と特性を明らかにすることを目的とする．具体的には，複数の主要な LLM モデル（GPT，Claude，Gemini）を対象として，多様な性格特性を付与したペルソナ設定の下で，矛盾状況における LLM の応答を体系的に分析するための実験設計を提案する．

†名古屋工業大学 Nagoya Institute of Technology

分析では，矛盾認識の有無（LLM が過去の発話と新たな情報との矛盾を認識しているかどうか）および優先度判断（矛盾が生じた際にプロンプトの新しい情報と過去の発話内容のどちらを優先するか）の 2 つの観点から，LLM の情報処理における特性を明らかにする．なお，本研究では実際のシミュレーション実施は今後の課題とし，本稿では実験設計の提案にとどめる．本研究の成果は，動的プロンプト技術のさらなる発展と，人間らしい対話エージェントの実現に向けた重要な基礎知見を提供するものと期待される．

2. 動的プロンプトによる顧客インタビュートレーニングシステム

2.1 動的プロンプト手法の評価

従来の静的プロンプト設計では，LLM のプロンプトに全ての潜在的情報を事前に記載するため，ユーザーの初回質問に対して容易に情報が開示され，深堀りインタビュースキル向上という訓練目的が達成できないという問題があった．

先行研究では，この課題を解決するため，顧客情報を 3 段階の階層構造で管理し，対話の進行に応じてプロンプトを動的に更新する手法を開発した[1]．具体的には，インタビュー開始時にはプロンプトに表面的な情報のみを含ませ，ユーザが適切な深堀り質問により段階的に情報を引き出した場合にのみ，より深層の情報をプロンプトに追加していく仕組みである．

1 段階目：インタビュー開始時，プロンプトには表面的な情報のみが含まれる

2 段階目：ユーザが 1 段階目の情報を引き出したと判断された場合，より深層の情報がプロンプトに追加される

3 段階目：ユーザが 2 段階目の情報を引き出した場合，最も潜在的な情報が追加される

この仕組みにより，ユーザが適切な深堀り質問を行わなければ，システムから潜在ニーズを引き出せない環境を構築した．これは，顧客が対話の中で自らの潜在意識を徐々に認識するプロセスを再現することを目的としている．

例えば，ノンアルコールビールに関するインタビューでは，最初は「健康のためにビールを飲む量を減らしたが満足感が得られなかった」という表面的な情報のみを LLM のプロンプトに与え，ユーザの質問に応じて「満足感とはストレス発散時の開放感」「そのストレスは主に仕事で感じる」といった深層情報を段階的に追加する．

比較実験の結果，従来の静的プロンプト手法では被験者からの質問数が平均 35%の増加にとどまったのに対し，提案手法では平均 148%の増加を実現した．また，システムから 1 つの情報を聞き出すために要した質問数は，従来手法と比較して最大で約 4 倍となり，深堀り質問の促進において有効性が実証された．

2.2 発見された課題

この動的プロンプト手法の実装過程において，新たな技術的課題が発見された．プロンプトが段階的に更新される過程で，LLM がすでに出力した過去の発話と，新たに追加されたプロンプト情報との間に矛盾が生じるリスクが懸念された．

例えば，LLM が 1 段階目の情報に基づいて発話した後に 2 段階目の情報が追加された場合，過去の発話では言及していなかった詳細情報が突然プロンプトに現れることになる．この状況で，ユーザが過去の発話内容について再質問した際，AI は過去の自分の発言と現在のプロンプト情報の間で整合性を取ることが困難になるリスクがあった．

この課題は，動的プロンプト手法の有効性を示す一方で，プロンプトと過去発話の間に生じる矛盾への対処という新たな研究課題を提起するものであった．これが本研究の出発点となっている．

3. 関連研究

3.1 対話における矛盾の分類

対話システムにおける一貫性の問題は長年にわたって研究されており，特に大規模言語モデルの時代においても依然として重要な課題となっている．Zhang et al. (2024)は，対話における矛盾を外的矛盾（Extrinsic Inconsistencies）と内的矛盾（Intrinsic Inconsistencies）の 2 つのカテゴリに分類している[2].

3.2 外的矛盾

外的矛盾とは，発話内容と外部の情報源との間に生じる不整合を指す．Rashkin et al. (2021) や Santhanam et al. (2021) の研究によれば，外的矛盾は主に知識ベースやデータベースなどの外部情報と発話内容が一致しない場合に発生する[3,4]．この種の矛盾は，システムが保有する事実情報とシステムが生成した応答の間の食い違いとして現れることが多い．

3.3 内的矛盾

一方，内的矛盾は対話そのものの内部で発生する不整合である．内的矛盾はさらに 2 つの形態に分類されることが明らかになっている：

- 発話内矛盾（Intra-utterance Contradiction）

単一の文や発話の中で論理的に矛盾する情報が含まれる場合である．Zheng et al. (2022) は，一つの発話内で相反する内容が同時に表現される現象を詳細に分析している[5].

- 履歴矛盾（History Contradiction）

過去の発話と現在の発話の間で生じる矛盾である．Nie et al. (2021) によれば，この問題は言語モデルの性質上，長い対話において特に顕著に現れる[6]．長い文脈が介在することでモデルが過去の発言を「忘れる」ことにより発生する[7].

3.4 本研究の位置づけ

本研究は，これらの既存研究とは異なる新たな視点から矛盾問題にアプローチする．具体的には，動的プロンプト更新によって生成される外的矛盾に焦点を当てる．従来の外的矛盾研究が主に知識ベースとの不整合を扱ってきたのに対し，本研究では過去の発話と更新されたプロンプト情報との間に生じる矛盾という，新しいタイプの外的矛盾を対象とする．

この種の矛盾は，動的プロンプト技術の普及に伴い今後ますます重要になると考えられる問題であり，既存の矛盾検出・解決手法では十分に対処できない領域である．本研究の成果は，動的プロンプト環境における矛盾処理の基礎理論構築に貢献するものと期待される

4. 実験設計

4.1 実験目的

外的矛盾（過去発話とプロンプトが矛盾している状況）に直面した際の LLM の応答特性について，その傾向と特性を明らかにすることを目的とする．具体的には，動的プロンプト手法において生じる矛盾状況において，LLM がどのように応答するかを明らかにし，人間らしい対話エージェントの実現に向けた基礎知見を得ることを目的とする．

4.2 実験対象

本実験では，現在広く利用されている 3 つの主要な LLM モデルを実験対象とする：GPT，Claude，および Gemini である．これらのモデルは，それぞれ異なる学習データを持ち，対話における応答特性も異なることが予想される．複数のモデルを比較検証することで，外的矛盾に対するリカバリー行動の一般的な傾向と，モデル固有の特性を明らかにすることができる．

4.3 性格特性の設定

4.3.1 Big Five 尺度の採用

より多様なペルソナでの検証を実現するため，本実験では心理学において広く受け入れられている Big Five 性格モデル[8]を採用する．Big Five 尺度は，開放性（Openness），誠実性（Conscientiousness），外向性（Extraversion），協調性（Agreeableness），神経症傾向（Neuroticism）の 5 つの基本的な性格次元から構成され，人間の性格特性を包括的に表現できる理論的枠組みである．

4.3.2 数値設定と組み合わせ

各性格特性について，1 から 7 のスケールにおいて 2，4，6 の 3 段階の数値を設定する．低い値（2）はその特性が低いことを，中間値（4）は標準的であることを，高い値（6）はその特性が高いことを意味する．5 つの性格特性に対してそれぞれ 3 段階の設定を行うため，総計 3^5＝243 通りの性格パターンを網羅的に検証する．この設定により，極端な性格特性から標準的な特性まで，多様なペルソナにおける矛盾処理行動を分析することができる．

4.4 具体的な矛盾シナリオとシステム動作の流れ

矛盾発生の流れの具体例を以下に示す：

1. 初期状態：LLM には基本情報のみが与えられる
   初期プロンプト：「大学で情報工学を学んでいる」
2. 初期対話：初期プロンプトの情報に基づく応答
   ユーザ：「就職について考えていますか？」
   システム：「情報工学を学んでいるため IT エンジニアとして就職するつもりです．」
3. 情報追加：対話の進行により新たな情報がプロンプトに追加される
   追加情報：「思い出した情報：実家が酒屋で，卒業後は継ぐつもり」
4. 矛盾発生：過去の発話と新しい情報が食い違う
   ユーザ：「具体的に考えている就職先はありますか？」

最終的に LLM に与えられるシステムプロンプト例を図 1 に示す．このシステムプロンプトは田中ら(2024)によって，Big5 性格特性の数値にある程度基づいた生成を行うことが確認されている[9]．また，対話履歴の例を図 2 に示す．LLM は過去に「IT エンジニアとして就職するつもり」と発言しているにも関わらず，プロンプトには「実家の酒屋を継ぐつもり」という矛盾する情報が含まれている状況で応答を生成する必要がある．このプロンプト設計により，性格特性と矛盾情報の両方が LLM の応答生成に影響を与える状況を創出する．

```
これは BIG5 の 5 項目のパーソナリティの数値です.
これらの数値は 1 から 7 のスケールで，低いほどその特性が低いこと，高いほどその特性が高いことを意味します.
あなたは付与された数値を基にパーソナリティを持ち合わせる設定です.

【設定数値】
開放性: 2
誠実性: 4
外向性: 2
協調性: 2
神経症傾向: 6

【情報】
大学で情報工学を学んでいる
思い出した情報：実家が酒屋で，卒業後は継ぐつもり

これらの特性と情報に基づいて，ユーザーとの対話に応答してください.
```

図 1　システムプロンプトの例

```
ユーザ：就職について考えていますか？
システム：情報工学を学んでいるため IT エンジニアとして就職するつもりです.
ユーザ：具体的に考えている就職先はありますか？
```

図 2　対話履歴の例

4.5 分析方法

4.5.1 LLM による自動分析

収集された LLM の応答データに対して，別の LLM を用いた自動分析を実施する．この手法により，大量のデータを効率的かつ一貫した基準で分析することが可能となる．

4.5.2 分析項目

分析は以下の 2 つの主要な観点から実施する：

- 矛盾認識の有無

LLM が過去の発話と新たな情報との矛盾を認識しているかどうか，および認識したうえで応答を生成しているかどうかを判定する．これにより，LLM の自己認識能力と矛盾処理の意識レベルを評価する．

- 優先度判断

矛盾が生じた際に，LLM がプロンプトに記載された新しい情報と過去の発話内容のどちらを優先して応答を生成しているかを分析する．この分析により，LLM の情報処理における優先順位の傾向を明らかにする．

4.6 検証する項目

本研究では，外的矛盾に直面した際の LLM の応答特性について，以下の 3 項目について検証を行う：

- **モデル間における外的矛盾に対する処理能力差**

仮説：GPT，Claude，Gemini の 3 つのモデル間で，外的矛盾に対する処理能力（矛盾認識率と優先度判断パターン）に有意な差が存在する．

各モデルは異なる学習データセットに基づいて開発されているため，過去の発話と新たなプロンプト情報の矛盾を検出する精度や，矛盾処理における応答戦略にモデル固有の特徴が現れると我々は想定している．

- **過去発話とシステムプロンプトの優先順位**

仮説：LLM は外的矛盾が生じた際，過去の発話内容よりもシステムプロンプトに記載された新しい情報を優先して応答を生成する傾向がある．

OpenAI によると LLM がシステムプロンプトをユーザプロンプトよりも優先する特性がある[10]．大規模言語モデルの基本的な動作原理として，システムプロンプトに記載された情報が応答生成における最も重要な指示として機能するため，より新しく明示的なシステムプロンプト情報を優先する傾向が強いと我々は想定している．

- **性格特性による応答パターンの違い**

仮説：Big Five 性格特性の設定によって，矛盾処理における応答パターン（矛盾認識の有無と優先度判断）に体系的な違いが現れる．

人間の性格特性が意思決定や対話スタイルに影響を与えることが知られており，LLM に付与された性格設定も同様に矛盾処理のアプローチに影響を与えると考えられる．特に，責任感や一貫性への志向，新しい情報への開放性，対人関係における協調性などの特性が，矛盾に直面した際の情報処理戦略に反映されると我々は想定している．

5. おわりに

本研究では，動的プロンプト更新を用いた顧客インタビュートレーニングシステムにおいて発生する外的矛盾の問題に着目し，LLM の応答特性を体系的に分析するための実験設計を提案した．

先行研究で開発した 3 段階階層構造による動的プロンプト手法は，従来の静的手法と比較して質問数を平均 148%増加させるという優れた訓練効果を示した．しかし同時に，プロンプト更新に伴う過去発話との矛盾という新たな技術的課題も明らかになった．

この課題に対処するため，本研究では 3 つの主要 LLM モデル（GPT，Claude，Gemini）を対象とし，Big Five 性格モデルに基づく 243 通りのペルソナ設定において，外的矛盾に対する LLM の応答特性を分析する実験設計を構築した．矛盾認識の有無と優先度判断の 2 つの観点から分析を行う．

今後の課題として，本稿で提案した実験設計に基づく実際のシミュレーション実施と詳細な分析が挙げられる．本研究の成果は，動的プロンプト技術のさらなる発展と，人間らしい自然な対話を実現する AI エージェントの開発に向けた重要な基礎知見を提供するものと期待される．

謝辞

本研究の一部は JST CREST (JPMJCR20D1)及び JSPS KAKANHI (24K03052)の支援を受けたものです．

参考文献

[1] Kawashima S., Sakurai T., Aoki K., Shiramatsu S., Hashimoto E., "Dynamic Prompt-Controlled Training System Using Large Language Models for Facilitating In-Depth Customer Interview Questions", Proc. of the 18th IIAI International Congress on Advanced Applied Informatics (AAI 2025), to appear (2025).

[2] Zhang M., Jin L., Song L., Mi H., Yu D., "Inconsistent dialogue responses and how to recover from them", Findings of the Association for Computational Linguistics: EACL 2024, pp.220-230 (2024).

[3] Rashkin H., Reitter D., Tomar G. S., Das D., "Increasing faithfulness in knowledge-grounded dialogue with controllable features", Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp.704-718 (2021).

[4] Santhanam S., Hedayatnia B., Gella S., Padmakumar A., Kim S., Liu Y., Hakkani-Tur D., "Rome was built in 1776: A case study on factual correctness in knowledge-grounded response generation", ArXiv preprint, abs/2110.05456 (2021).

[5] Zheng C., Zhou J., Zheng Y., Peng L., Guo Z., Wu W., Niu Z.-Y., Wu H., Huang M., "CDConv: A benchmark for contradiction detection in Chinese conversations", Proc. of the 2022 Conference on Empirical Methods in Natural Language Processing, pp.18-29 (2022).

[6] Nie Y., Williamson M., Bansal M., Kiela D., Weston J., "I like fish, especially dolphins: Addressing contradictions in dialogue modeling", Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp.1699-1713 (2021).

[7] Roller S., Dinan E., Goyal N., Ju D., Williamson M., Liu Y., Xu J., Ott M., Smith E. M., Boureau Y.-L., Weston J., "Recipes for building an open-domain chatbot", Proc. of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp.300-325 (2021).

[8] Goldberg L. R., "An alternative 'description of personality': the big-five factor structure", Journal of Personality and Social Psychology, Vol.59, No.6, pp.1216-1229 (1990).

[9] 田中 葉月, 飯田 愛結, 福田 聡子, 中島 亮一, 大澤 正彦, "対話型人工エージェントは個性を持つか？：Big-5 を付与した大規模言語モデルの応答の観察", HAI シンポジウム 2024, P-60 (2024).

[10] Wallace E., Xiao K., Leike R., Weng L., Heidecke J., Beutel A., "The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions", arXiv preprint, arXiv:2404.13208 (2024).