公開日：2023/12/08（初版）
作成者： mercari R4D

LLMを用いたプロダクト開発をスピーディーに行うためのガイドライン

1. はじめに

このガイドラインのスコープ

本ガイドラインは、LLMの既存モデルをプロダクトに適用する際に、開発メンバーが参照するためのものです。
本ガイドラインでは、OpenAI社が提供するAPIの利用を想定しています。
【補足：LLMのモデルや利用方法（API連携か否か）により機械学習の有無、学習用データの許諾の有無等、法的リスクが異なるため、他のLLMや利用方法を対象とする場合には、個別に検討が必要となります。】

2. 注意事項の概要

本ガイドラインで定める注意事項等の概要は以下のとおりです。各注意事項の詳細は「3. リスク対策詳細」もあわせてご確認ください。

（1）ユースケースと注意事項

(i):「LLMの直接的な利用者が従業者である(社外には、”LLMにより生成したコンテンツ”を提供する)場合」のリスク対策と手続き

LLMの直接的な利用者が従業者である場合の例:
- LLMからの出力を社内メンバーしか利用しない
- LLMからの出力に基づいて生成した生成物をお客さまに提供する
- LLMからの出力を社内で事前に確認したうえで、お客さまに提供する
- お客さまからの入力に基づかずにLLMが出力を生成する

このケースにおいて、LLMからの出力をお客さまに提供する場合には、事前に提供物に意図しない情報が含まれていないかを確認し、精査してください。具体的には以下のような情報が含まれていないかの確認が必要です

- 有害な情報(犯罪を助長する情報、性的・差別的・侮辱的な表現等)
- 社内の機密情報
- 他人の個人情報(元の個人情報の利用目的の範囲内か否かの確認を含む)
- 第三者との間の秘密保持契約に基づき取得した情報その他第三者の機密情報
- 法律等の制限によって禁止されている行為(特定の金融商品への勧誘、法律相談、医療相談等)
- 他人の権利(著作権、商標権、肖像権、パブリシティ権等)を侵害する情報
    - ※特に、特定のアーティストや作品名等を指定してその作風を模した画像等を出力させた場合、著作権の侵害と認定されるおそれが高まります。
- 事実に基づかない誤った情報

(※これらの意図しない情報が含まれるリスクについて、併せて「3. リスク対策詳細>②画像、音楽、物語等の著作物になり得るものを出力する場合は、知財チームに相談する、③意図しない情報の出力が抑制されている」をご参照ください)

なお、このケースにおける利用可能な情報の種類については、「(2):promptに含むことができる情報」をご確認ください。

(ii):「LLMの直接的な利用者がお客さまである(LLMを用いたサービスを提供する)場合」のリスク対策と手続き

LLMの出力をお客さまがリアルタイムに直接受け取る場合には、必要なリスク対策は、お客さまにChat形式等、入力の自由度の高いインターフェースを提供するかどうかで分類されます。以下にそれぞれの場合の対策と、手続きについて示します。
なお、利用可能な情報の種類については、「(2):promptに含むことができる情報」(注:本項目においては、API連携時において社内データベース

と連携されるデータの範囲を指します。）をご確認ください。

入力形式	必要なリスク対策
入力の自由度の低いインターフェースの場合	①お客さまがLLMを利用することのリスクについて、把握した上で利用できるよう配慮されている （例） ● 利用開始前にLLMを利用することの注意事項を提示する ● LLM利用に関するリスクが表示されるUIになっている ②画像、音楽、物語等の著作物になり得るものを出力する場合は、知財チームに相談する ③意図しない情報の出力が抑制されている （例） ● 有害な情報（犯罪を助長する情報、性的・差別的・侮辱的な表現等）や他人の個人情報、社内の機密情報、第三者との間の秘密保持契約に基づき取得した情報その他第三者の機密情報、法律等の制限によって禁止されている行為（特定の金融商品への勧誘、法律相談、医療相談等）等が出力されないように設計されている （コンテンツフィルタリング（Moderation API）、出力のテンプレ化、等） ● 本来参考にしてほしくない情報（誤った情報等）に基づいて回答を出力しないよう設計されている （マイクロソフト Azure Prompt Flow (Groundness)のようなツールでのチェック等） ④prompt injectionやLLMにおけるよくある脆弱性への対策が取られている ⑤API提供元のサービスが不安定になった場合の対応が考慮された設計になっている ⑥入出力のログが取得されている

Chat形式等、入力の自由度の高いインターフェースの場合	上記①～⑥に加え、 ⑦お客さまがAIと対話していることが理解できるUIになっている ⑧意図しない情報の入力が抑制されている（入力の自由度が上がることにより、③の対策を回避されるリスクがあることを意図した設計になっている） （入力に対してもバリデーションを入れる、特定の質問以外は捨てる、等）

（2）promptに含むことができる情報（（1）(i),(ii)双方の場合）

promptに含まれる情報の種類 ※回答生成の補助情報として**prompt**で与えることを含む	
Publicな情報のみ	制限なし
機密レベルが**Confidential**以下の情報（＝全社に公開されている情報） （例） • ワーディングルール • マニュアル • 個人情報を匿名加工した統計情報	利用する情報の種類、件数、利用態様（作成するプロダクトの概要等）を、関連チームに共有し、事前に相談してください ※なお、お客様に関連する情報を利用する場合には、利用目的が同意取得済みの目的の範囲内である必要があります。 ※一見すると、個人情報には思えない情報が意外と個人情報である場合があります。機密情報を利用する場合には、一度、利用したい情報に個人情報が本当に含まれていないか、関連チームに確認をとるようお願いします。 【補足：入力した情報がAIサービスの提供事業者に学習目的で利用

	されない場合であっても、当該AIサービスの規約上、システム監査・保守等の目的で利用されることがある旨規定される場合があります。そのため、機密情報の入力を一律に制限する建付けとすることも考えられます。実際の運用に応じ、適宜ご修正ください。】
個人情報 （例） ● PII（氏名、住所等、それ単体で個人を識別できる情報）そのもの ● 信用情報（個人の返済能力等に関する情報） ● 他の情報と容易に照合することができ、それにより特定の個人を識別することができることとなるもの	利用を希望する場合は関連チームにご相談ください。 ※リスク対策の実施状況や情報の個別の種類によっては利用可能な場合があります。 【補足：入力した情報がAIサービスの提供事業者に学習目的で利用されない場合であっても、当該AIサービスの規約上、システム監査・保守等の目的で利用されることがある旨規定される場合があります。そのため、機密情報の入力を一律に制限する建付けとすることも考えられます。実際の運用に応じ、適宜ご修正ください。】
Strictly Confidential、**Top Secret**に相当する情報 （例） ● 秘匿性の高いソースコード ● 他社の秘密情報 ● アクセス権が一部のメンバーに限定されている情報 ● 他者が権利を有する著作物 ● 特許化を検討している発明 ● その他、これらに準ずる漏洩リスクが特に高い情報 パスワードや**API**キー等、個人で秘密として管理すべき情報の利用	

（3）利用規約の確認及び遵守（（1）(i),(ii)双方の場合）

AIサービスを利用するにあたっては、通常、当該AIサービスの提供事業者が定める利用規約を確認の上、同意することが必要となります。特に、

以下の事項についてはAIサービスを利用の前提として確認する必要が高いため、よくご確認ください。

- 商用利用の可否【補足：ChatGPTはこれに該当しませんが、生成結果の商用利用が不可とされている場合には、お客さまに生成結果を提供することは利用規約違反になるおそれがありますのでご留意ください。】
- 入力情報の取扱い態様（入力した情報が当該AIサービスの提供事業者が当該AIの学習に利用したり、サービスの開発に利用されるか否か）
- 生成結果がAIによって生成されたものであることを示す表示を付す義務の有無

（4）ユースケースに基づくリスク分類

リスクのレベル	内容	インシデントが発生する可能性があるケース	リスクがコントロール可能なケース
高	個人情報の流出 例えば、ユーザAへの回答を生成するにあたり、ユーザA,B,Cのサービス利用履歴等の個人情報が補助情報としてpromptに入力されるよう設計した場合、ユーザAに、ユーザB,Cの利用履歴の情報が出力されるおそれがある。	LLMに個人情報を与える場合	● LLMに個人情報を与えない ● Promptに当該お客様の情報のみが入力されるようにし、お客さま本人にお客さま自身の情報を出力する場合 ○ リスクを許容することが個別に判断されている
高	禁止行為の出力 ● 法律等の制限によって禁止されている行為（特定の金融商品への勧誘、法律相談、医療相談等）の出力	Chat形式でお客さまと自由に会話させる場合 ※入出力のバリデーション、LLM自体の機能でほぼ0近くに低減可能	● LLMの出力を社内メンバーしか見ない ● お客さまからの入力に自由度がないUI（Chat形式以外のUI） ● Chat形式でお客さまに出力する場合 ○ 禁止行為となる情報の出力がされるようなプロンプトの入力及び禁止行為となる情報の出力を防止する措置・お客さまへの注意喚起が適切にされている
中～高	重大な機密情報の漏洩	Chat形式でお客さまと自由に会話させ、かつ機密情報（Strictly	● LLMに重大な機密情報を与えない

	prompt injection等の攻撃により、本来秘密として管理しておくべき重要な機密情報（Strictly Confidential相当）が漏洩する	Confidential相当）をLLMに与える場合	● 具体的な社名、商品名、人名、金額等が特定できないように情報を加工して入力する ● お客さまからの入力に自由度がないUI（Chat形式以外のUI） ●
中	<u>他人の著作権等の権利を侵害</u> LLMからの生成物が、既存の著作物と同一・類似している場合は、当該生成物を利用（複製や配信等）する行為が著作権侵害に該当する可能性がある。	LLMに画像や写真等を生成させる場合	● LLMに画像等を出力させない ● LLMに学習されているコンテンツがパブリックドメイン又は適切に権利処理が行われているもののみである【補足：権利侵害に関するトラブルを可及的に回避する観点からは、学習データが権利処理されているAIサービスのみを利用可能とするという運用も考えられます。ChatGPTはこれに該当しないものの、例えば、Adobe社が提供するFireflyがこれに該当します（下記①、②及び③9.参照）。 1 よくある質問（webアプリ版） 2 よくある質問（エンタープライズ版） 3 Firefly Legal FAQs – Enterprise Customers】 ● 権利侵害を構成するプロンプトの入力及び権利侵害を構成するコンテンツが生成されることを防止する措置が適切に行われている
低～中	<u>有害な情報の出力</u> ● 性的な表現 ● 犯罪を助長する表現 ※利用するLLMの学習データにこれらの情報が含まれていた場合、LLMが会話の流れによってはこのような情報を出力してしまう	Chat形式でお客さまと自由に会話させる場合 ※入出力のバリデーション、LLM自体の機能でほぼ0近くに低減可能	● LLMの出力を社内メンバーしか見ない ● お客さまからの入力に自由度がないUI（Chat形式以外のUI） ● Chat形式でお客さまに出力する場合 ○ 有害情報を生成するようなプロンプトの入力及び有害情報を含むコンテンツが生成されることを防止する措置・お客さまへの注意喚起が適切にされている
低	<u>誤情報の出力</u> 大規模言語モデル（LLM）の原理は、「ある単語の次に	Chat形式でお客さまと自由に会話させる場合	● お客さまからの入力に自由度がないUI（Chat

	用いられる可能性が確率的に最も高い単語」を出力することで、もっともらしい文章を作成していくものであるため、出力された文章が事実に反する場合又は正しくない場合でもそれが分かりづらい性質がある。LLM性質に依存するリスク。	※発生確率は完全に0にはならないがリスクは小さい	形式以外のUI） ● Chat形式でお客さまに出力する場合 ○ 誤情報を生成するようなプロンプトの入力及び誤情報を含むコンテンツが生成されることを防止する措置・お客さまへの注意喚起が適切にされている

3. リスク対策の詳細

関連するリスク

Risk Category	Risk 概要	例
害悪	AIによる犯罪や不適切行為の助長	自社が提供したchatにおいて、犯罪を助長するアドバイスが行われたり、性的な表現の回答が出力される
バイアス 差別 排除	学習データに起因して発生するリスク ● バイアス：過去のレスポンスや学習データからバイアスを増幅 ● 差別：過去のデータに含まれていた差別や偏見をAIが学習し、アウトプットに反映してしまう ● 排除：特定の集団（人種、ジェンダー、移住地など）の人々に関する情報が欠落していることによる疎外	バイアス等を反映した結果、不当に選択肢が狭められたり、不当に低評価を与えられうる 政治扇動や、宗教対立を助長する

自動化バイアス	自動化されたシステムや技術に、欠陥や信頼性がないにもかかわらず、人間が過度に依存してしまう	LLMの出力を受け取った利用者が、その情報の正誤に関わらず正しいと信じて発信してしまう
誤情報 幻覚	幻覚（真実であるように生成されるリスク）誤った情報、一貫性のない情報を出力してしまう	LLMの出力を受け取った利用者が、その誤った情報を正しいと信じて発信してしまう 幻覚、文脈のない回答、一貫性のなさなどにより混乱を与えうる、また誤情報に基づく意思決定を生む可能性
プライバシー	機微データやプライバシーに関する情報や画像が生成・漏洩し、プライバシーが侵害される	個人情報の漏えい、それを第三者に操作されうる 利用者の意図に関わらず出力に他人のプライバシーを侵害する情報が含まれる
セキュリティ	prompt injectionなどLLMにおける脆弱性に対する攻撃手法が知られている 攻撃手法の一例 • OWASP Top 10 List for Large Language Models version 0.1	攻撃により、回答の補助情報としてpromptに入力したものの本来出力させず秘密として管理されるべき情報が漏洩してしまう
他人の権利の侵害	AIで生成した文章や画像による著作権・肖像権の侵害（inputの著作権、outputの著作権）	有名人の写真を用いて出品物の着画像を生成した場合、パブリシティ権を侵害する可能性がある

リスク対策詳細

①お客さまがLLMを利用することのリスクについて、把握した上で利用できるよう配慮されている
⑦利用者がAIと対話していることが視認できるUIになっている

対象リスク:

- バイアス /差別 / 排除
- 自動化バイアス
- 誤情報 / 幻覚
- 他人の権利の侵害

（説明）

LLMの出力した情報には上記のようなリスクがあります。これらのリスクはそもそもの学習データや、LLMの特性に起因して発生するため、100%取り除くことは難しいです。

そのため、お客さま自身に、LLMにはこのようなリスクがあることを理解し、自衛していただくことが重要です。そのためにも、(i)直接的な利用者が従業者である(LLMにより生成したコンテンツを社外に提供する)場合において、生成結果の真実性について、可能な限り自ら裏付けの有無の確認及び校閲を実施することに加え、(ii)直接的な利用者がお客さまである(LLMそのものをサービスとして提供する)場合において、利用規約上、生成された文章に誤りが含まれる可能性があり、真実性を保証するものではないことや、お客さまFによる目的外の利用、悪意のある利用等により生じたトラブル等に対する免責事項を明記することが肝要です。

②出力に画像、音楽、物語等の著作物になり得るものは含む場合は知財チームに相談する

対象リスク:

- 他人の権利の侵害

（説明）

LLMからの生成物が、既存の著作物と同一・類似している場合は、当該生成行為又は当該生成物を利用（複製や配信等）する行為が著作権侵害に該当する可能性があります。特に、特定のアーティストや作品名等を指定してその作風を模した画像等を出力させた場合、著作権の侵害と認定されるおそれが高まります。
また、第三者が商標権や意匠権を有するロゴ又はデザイン等を入力すると、第三者の登録商標又は意匠と同一又は類似の出力結果が生成される可能性があり、当該生成結果の利用行為が第三者の商標権や意匠権侵害に該当する可能性があります。したがって、これらの著作物等の入力を行わないことが望ましいです。
さらに、有名人の氏名、芸名、肖像、音声等のパブリシティと同一又は類似の生成結果を利用する行為（有名人の写真を用いて出品物の着画像を生成させた場合等）には肖像権やパブリシティ権を侵害する可能性があります。
画像や動画等のコンテンツを出力するように設計する必要がある場合には、以下を実施のうえ、知財チームに相談してください

- 他人の権利を侵害しない、あるいは侵害した場合検知できる仕組みを設ける【補足：例えば、Google画像検索等により類似する画像がないかを事後的に検証することのほか、事前の仕組みとしては、著作物や他人の著名な著作物に類似したものを生成することを意図したプロンプトが入力されたことを検知する仕組み等が考えられます。ただし、この場合であっても侵害リスクを完全に排除できるものではありませんので、ご留意ください。】
- (i)直接的な利用者が従業者である（LLMにより生成したコンテンツを社外に提供する）場合、著作物については、ウェブ検索エンジン上で類似する既存著作物等がないかを確認すること、商標や意匠については、担当者自身によるJ-PlatPatを用いた調査のほか、適宜、顧問弁理士への相談も検討する
- (ii)直接的な利用者がお客さまである（LLMそのものをサービスとして提供する）場合、生成物の権利の非侵害性について保証しないことを利用規約上明記する

③意図しない情報の出力が抑制されている
⑧意図しない情報の入力が抑制されている（入力の自由度が上がることにより、③の対策をかいくぐられるリスクがあることを意図した設計になっている

対象リスク：

- 害悪
- 誤情報

- 幻覚
- プライバシー

（説明）

LLMには、犯罪を助長する情報や性的な表現、他人のプライバシーを侵害するような情報等の、開発者が意図していない情報を出力してしまうリスクがあります

- 例1：ユーザAへの回答を生成するにあたり、ユーザA, B, Cのサービスの利用履歴等の個人情報が補助情報としてpromptに入力されるよう設計した場合、ユーザAに、ユーザB,Cの利用履歴の情報が出力されるおそれがある。
- 例2：CSお問い合わせbot。CSマニュアルを補助情報として与えて回答するよう設計していても、モデル作成時に学習データとして与えられた情報（例えばユーザがネット上に公開していたサービスの体験記等）や、Chatの履歴で得た情報をLLMが勝手に優先して回答を生成してしまい、結果、公式には認めていない回答を出力するおそれがある。

特に、Chat形式のインターフェースの場合、お客さまから想定外の入力をされることによって、開発者が意図しない出力をLLMがしてしまうリスクが上がります。このようなリスクを防ぐためには、（1）漏えいリスクの高い情報をそもそもpromptに含めない、（2）お客さまからの入力の自由度を下げる（入力のテンプレ化、特定の質問以外は捨てる等）、（3）コンテンツフィルタリング等で出力側で制限をかける、等の仕組みを設ける必要があります。

また、このような仕組みが正しく機能しているかリリース前に必ず検証するようにしてください。

④prompt injectionやLLMにおけるよくある脆弱性への対策が取られている

対象リスク：

- プライバシー
- セキュリティ対策

（説明）

- OWASPのガイドライン等を参照し、適切なリスク対策を実施する
    - OWASP Top 10 List for Large Language Models version 0.1
- プロダクトリリース後も新たな攻撃手法が出現しないか定期的にモニタリングを行い、随時対応を行ってください。なお、当該モニタリングのレビュー体制は、「(ii):「LLMの出力をお客さまがリアルタイムに直接受け取る場合」のリスク対策と手続き」と同様です

⑥入出力のログが取得されている

対象リスク:

（説明）

- LLMは新しい技術であり、まだすべてのリスクが顕在化されているとは言えません。そのため、入出力のログを保存・レビューし、品質の改良に務めることが重要です