汎用 AI 実務規範第 3 ドラフト

汎用 AI モデルのプロバイダによるコミットメント

透明性セクション

透明性部会長および副部会長による序文

実務規範の透明性の項では、第 53 条(1)(a)および(b)、ならびに AI 法附属書 XI および XII に規定される透明性義務を遵守するために、 署名者が取るべき 3 つの施策について記述している。

この第 3 ドラフトでは、I.1.1 項に含まれる約束の履行を合理化し、署名者の遵守を促進するため、必要な情報を一箇所で簡単に文書化できるよう、使いやすいモデル文書様式を盛り込んだ。

本様式は，各項目について，それが川下プロバイダ，AI 事務局又は各国所轄当局のいずれを対象 とするものであるかを明確に示している。川下プロバイダ向けの情報は，川下プロバイダが積極的に利用できるようにすべきであるが， AI 事務局又は国家管轄当局向けの情報は，AI 事務局からの要請に基づき，又は国家管轄当局から AI 事務局 への要請に基づき，職権で利用できるようにする。このような要請には、要請の法的根拠と目的が明記され、要請時に AI 事務局が AI 法に基づく任務を遂行するため、または各国所轄官庁が要請時に AI 法に基づく監督任務を遂行するため、特にシステムのプロバイダがモデルのプロバイダとは異なる汎用 AI モデル上に構築された高リスク AI システムのコンプライアンスを評価するために、厳密に必要なフォームの項目のみが対象となる。

最後に、AI 法第 78 条に従い、モデル・文書化・フォームに含まれる情報の取得者は、取得した情報、特に知的財産権、業務上の機密情報または企業秘密の秘密を尊重し、取得した情報のセキュリティと機密性を保護するために適切かつ効果的なサイバーセキュリティ対策を講じる義務がある。

ヌリア・オリバー

第 1 作業部会 共同議長

リシ・ボマサニ

第 1 作業部会

透明性セクションの前文

以下の通り：

a) なぜなら、提供されるモデルは、様々な下流システムの基盤となる可能性があり、多くの場合、下流プロバイダは、そのようなモデルの製品への統合を可能にし、AI 法に基づく義務を果たすために、モデルとその機能について重要な理解を必要としているからである（AI 法 101 項参照）。

b) 署名者は、モデルの修正や微調整の場合、比例性を守るため、プロバイダの義務は修正や微調整に限定されるべきであると認識する（AI 法 109 条前文参照）。

コミットメント I.1. 文書化

法律文書：第 53 条第 1 項（a）、第 53 条第 1 項（b）、第 53 条第 2 項、第 53 条第 7 項、AI 法附属書 XI および XII

AI 法第 53 条 1 項(a)および(b)の義務を果たすため、署名者は、I.1 項に従ってモデル文書を作成し、最新の状態に保つ。.1、施策 I.1 に従い、汎用 AI モデルを AI システムに統合しようとする AI システムのプロバイダ（以下、川下プロバイダ）、および要請に応じて AI 事務局に関連情報を提供すること（特に、システムのプロバイダがモデルのプロバイダと異なる場合、汎用 AI モデル上に構築された高リスク AI システムのコンプライアンスを評価するため、AI 法に基づく監督業務の遂行に厳密に必要な場合、AI 事務局に要請があれば、各国所轄官庁に代わって AI 事務局に提供することも可能である[1]）。.2 に従い、また、施策 I.1.3 に従い、文書化された情報の品質、安全性及び完全性を確保する。本施策は、AI 法第 53 条第 2 項に定める条件を満たすオープンソースの AI モデルのプロバイダには適用されないが、当該モデルがシステミック・リスクを有する汎用 AI モデルである場合はこの限りではない。

施策 I.1.1. モデル文書を作成し、最新の状態に保つ

署名者は、汎用 AI モデルを上市する際、以下の「モデル文書様式」で言及されているすべての情報を含む「汎用 AI モデルに関する情報および文書」と題する文書（以下、モデル文書）を作成することを約束する。

署名者は、比較可能な、検証可能な文書を作成するため、AI 法第 53 条 5 項に従って採択された、測定・計算方法を詳細に規定する委任法に従い、計算資源およびエネルギー消費量セクションで要求される情報を報告することを約束する。

モデル文書に含まれる情報に関連する変更があった場合、署名者は、新情報を反映するようモデル文書を更新する一方、モデル文書の旧バージョンを、そのモデルが市場から撤去されてから 10 年間保管することを確約する。

施策 I.1.2. 関連情報のプロバイダ

署名者は、汎用 AI モデルを上市する際、AI 事務局および川下プロバイダがモデル文書に含まれる関連情報の閲覧を要求するための連絡先を、ウェブサイトを通じて、またはウェブサイトを持たない場合は他の手段を通じて、一般に開示することを約束する。

AI 法第 91 条または第 75 条第 3 項に基づき、AI 事務局が AI 法に基づく業務を遂行するため、あるいは各国主管庁が AI 法に基づく監督業務を遂行するために厳密に必要なモデル文書の 1 つ以上の要素について、AI 事務局から要求があった場合、

[1] AI 法 75 条 1 項および 3 項、AI 法 88 条 2 項を参照のこと。

特に、システムのプロバイダがモデルのプロバイダと異なる場合、汎用 AI モデル上に構築されたハイリスク AI シ ステムのコンプライアンスをアセスメントするため[1] 、署名者は、AI 法第 53 条第 7 項および第 78 条に規定される守秘義務 および条件に従い、最新のモデル文書に含まれる関連要素、または必要な追加情報を提供することを約束する。

署名者は、第 53 条 7 項および AI 法第 78 条に規定される守秘義務および条件に従い、最新のモデル文書に含まれ る、下流プロバイダ向けの情報を下流プロバイダに提供することを約束する。さらに、同じ機密保護施策と条件に従い、署名者は、川下プロバイダが汎用 AI モデルの能力と限界を十分に理解し、AI 法に基づく義務を遵守するために必要な追加情報を提供することを約束する。

署名者は、上記のすべての行動を適時にとることを約束する。

署名者は、透明性を促進するため、文書化された情報の全部または一部を公開できるかどうかを検討することが奨励される。この情報の一部は、AI 事務局が提供するテンプレート（ ）により決定される AI 法第 53 条 1 項（d）に基づき、プロバイダが一般に公開しなければならない研修内容の要約の一部として、要約された形で要求される場合もある。

施策 I.1.3. 情報の品質、完全性、およびセキュリティの確保

署名者は、文書化された情報の品質と完全性を管理し、AI 法の義務遵守の証拠として保持し、意図しない改ざんから保護 することを約束する。情報および記録の作成、更新、品質と安全性の管理に関して、署名者は、確立された手順と技術標準に従う よう奨励される。

モデル・文書化・フォーム

以下は、モデル文書フォームの静的で編集不可能なバージョンである。このバージョンでは、入力フィールドに記入することはできず、フォームの下部にあるボタンは、川下プロバイダ向けの情報のみを含むフォームのバージョンを生成するためのものであるが、機能しない。本規範の最終ドラフトでは、このフォームは完全に双方向で編集可能となる。

[1]AI 法 75 条 1 項、3 項、88 条 2 項参照

モデル・文書化・フォーム

本書式には，施策 *1.1* の一部として文書化されるすべての情報が含まれる。右側の×印は、文書化された情報が *AI* 事務局（*AIO*）、各国所轄庁（*NCA*）、または川下プロバイダ（*DP*）のいずれを対象としているかを示している。

は、汎用 *AI* モデルを *AI* システムに統合することを意図している。*DP* 向けの情報は、*DP* が利用できるようにすべきである。

AIO または *NCA* 向けの情報は、*AIO* からの要請に基づき、職権で、または *NCA* から *AIO* への要請に基づいてのみ、積極的に提供される。このような要請には、要請の法的根拠と目的が明記され、「書式」の項目のみが対象となる。

特に、システムのプロバイダとモデルのプロバイダが異なる場合に、汎用 *AI* モデルに基づいて構築された高リスク *AI* システムのコンプライアンスを評価するために、*AIO* が要請時に *AI* 法に基づく業務を遂行するため、または *NCA* が要請時に *AI* 法に基づく監督業務を遂行するために必要なもの。

AIO、NCA または DP と共有されるモデル文書様式の情報の要素は、第 78 条に定める秘密保持義 務および企業秘密保護に従って取り扱われるものとする。

文書の最終更新日　　**資格文書のバージョン番号**

一般情報			AIO	NCA	DPs
プロバイダの正式名称			☒	☒	☒
モデルファミリー	モデルのコレクションの識別子があれば、それを指定する（例：Liama		☒	☒	☒
モデル名・バージョン	モデル固有の識別子（例：Llama 3.1-405B）		☒	☒	☒
モデル認証	モデルの出所と認証を証明する証拠（バイナリが配布されている場合は安全なハッシュ、サービスの場合は URL のエンドポイントなど）。		☒	☒	☐
販売日		そのモデルが何らかの流通経路で最初に発売された日付。	☒	☒	☒
欧州市場導入日		欧州市場に投入された日付。	☒	☒	☒
モデルの依存関係	もしあれば、そのモデルが依存している他の汎用 AI モデルのリスト（例えば、llama-3.1-nemotron-70b のリストは[llama-3.1]となり、llama-3.1 のリストは空となる）。リストアップされた各モデルの依存関係について、関連するモデルドキュメントのコピーまたはリンクを含めるか、モデルドキュメントにアクセスできないことを示すこと。		☒	☒	☒

一般情報						AIO	NCA	DPs
モデルアーキテクチャー	モデル・アーキテクチャの一般的な説明（例：トランスフォーマ・アーキテクチャ）。[推奨 20 語]					☒	☒	☒
	モデルがシステミックリスクを有する汎用 AI モデルである場合、モデルアーキテクチャの詳細な説明を記載し、該当する場合は標準アーキテクチャから逸脱する箇所を特定すること。モデルがシステミックリスクを有する汎用 AI モデルでない場合、「N/A」と記入すること。					☒	☐	☐
モデルの設計仕様	モデルの主要な設計上の選択肢についての一般的な説明。その根拠や前提条件を含め、モデルがどのように設計されたかについての基本的な理解を提供する。 [推奨 100 語]					☒	☒	☐
						☒	☒	☒
入力モード	☐ **テキスト**	☐ **画像**	☐ **音声**	☐ **動画**	**その他があれば明記**	☐	☐	☒
選択した各様式について、最大入力サイズを記載するか、定義されていない場合は「N/A」と記入すること	最大サイズ	最大サイズ	最大サイズ	最大サイズ	最大サイズ	☒	☒	☒
出力モード	☐ **テキスト**	☐ **画像**	☐ **音声**	☐ **動画**	**その他があれば明記**	☐	☐	☒
選択した各様式について、最大出力サイズを記載するか、定義されていない場合は「N/A」と記入すること	最大サイズ	最大サイズ	最大サイズ	最大サイズ	最大サイズ	☒	☐	☐
						☐	☒	☒
総合モデルサイズ	モデルのパラメータの総数。少なくとも 2 桁の数字で記録する。例えば、7.3*10^10 パラメータ。							
パラメータの総数が属する範囲を選択する	☐ **1-500M**	☐ **500M-5B**	☐ **5B-15B**	☐ **15B-50B**				
	☐ **50B-100B**	☐ **100B-500B**	☐ **500B-1T**	☐ **1T>**				

流通とライセンスの方法		AIO	NCA	DPs
流通チャネル	モデルプロバイダが把握している範囲で、外部関係者がモデルにアクセスできるすべての配布チャネ ル（例えば、既存のソフトウェアスイートまたはエンタープライズ固有のソリューションを通じたエンタープライズ またはサブスクリプションベースのアクセス、API を通じたパブリックまたはサブスクリプションベースのアクセ ス、統合開発環境、デバイス固有のアプリケーションまたはファームウェア、オープンソースリポジト リを通じたパブリックまたはプロプライエタリなアクセス）のリスト。記載された各流通経路について、利用可能な場合、どのようにモデルにアクセスできる か、また、その経路を通じたモデルアクセスのレベル（例：ウェイトレベルアクセス、ブラックボッ クスアクセス）に関する情報へのリンクを含めること。	☒	☒	☒
ライセンス	モデルライセンスへのリンク（または、コピーをこの文書に添付する）または、存在しない場合はその旨を記載する。	☒	☒	☐
	モデルが下流のプロバイダに提供される可能性があるライセンスの種類またはカテゴリー。	☐	☐	☒
		☒	☒	☐

	追加の資産（例えば、訓練データ、データ処理コード、モデル訓練コード、モデル推論コード、モデル評価コード）の一覧（もしあれば）と、各資産へのアクセス方法、および使用に関連するライセンス（もしあれば）の説明。			

利用		AIO	NCA	DPs
利用規定	適用される利用規約へのリンクを提供すること（または、この文書にコピーを添付すること）、または利用規約が存在しないことを示すこと。	☒	☒	☒
使用目的	利用（i）プロバイダが意図する利用（生産性向上、翻訳、創造的コンテンツ生成、データ分析、データ視覚化、プログラミング支援、スケジュール管理、カスタマーサポート、自然言語処理など）の説明、または または（ii）プロバイダによって制限および/または禁止されている用途（EU 法または国際法によって禁止されている用途を超えたもの、AI 法第 5 条を含む）の両方の場合において、プロバイダが提供する使用説明書、利用規約、販促または販売資料および文書、ならびに技術文書に規定されている。(i)または(ii)を指定することが、モデルが提供されるライセンスの性質と矛盾する場合は、「該当なし」と記入できる。 [推奨 200 語]	☒	☒	☒
汎用 AI モデルを統合できる AI システムのタイプと性質：	(i)汎用 AI モデルを統合できる AI システムのタイプと性質、または(ii)汎用 AI モデルを統合すべきでない AI システムのタイプと性質、のリストまたは説明。例としては、自律システム、会話アシスタント、意思決定支援システム、創造的 AI システム、予測システム、サイバーセキュリティ、監視、または人間と AI の協働などが含まれる。 [300 語以内を推奨]	☐	☐	☒
モデル統合のための技術的手段：	汎用 AI モデルを AI システムに統合するために必要な技術的手段（使用説明書、インフラ、ツールなど）の一般的な説明。 [推奨 100 語]	☐	☐	☒
必要なハードウェア	該当する場合、モデルを使用するために必要なハードウェア（バージョンを含む）の説明。該当しない場合（API 経由で提供されるモデルなど）は、「該当なし」と入力する。[推奨 100 語]	☐	☐	☒
必要なソフトウェア	該当する場合、モデルを使用するために必要なソフトウェア（バージョンを含む）の説明。該当しない場合は、「該当なし」と記入する。[推奨 100 語]			

訓練プロセス		AIO	NCA	DPs
訓練プロセスの設計仕様	訓練プロセスに関わる主なステップや段階の一般的な説明、訓練方法や技術、主な設計上の選択、前提条件、モデルが最適化される対象など。例えば、「モデルはランダムに選択された重みで初期化され、2 段階で Adam オプティマイザーによる勾配ベースの最適化を用いて最適化される。まず、モデルは、クロスエントロピー損失を使用して、1 つのエポックのデータを処理しながら、大規模な事前学習コーパス上の次の単語を予測するように訓練される。次に、モデルは、人間の価値観と一致させ、ユーザーの指示への対応をより有用なものにするために、人間の好みに関するデータセットで 10 エポックの事後訓練を受ける。[推奨 200 語]	☒	☒	☐
異なるパラメータの関連性：	適用可能な場合、異なるパラメータの関連性。	☒	☒	☐
決定の根拠	モデル訓練における主要な設計上の選択がどのように、なぜなされたかの説明。 [推奨 200 語]	☒	☒	☐

訓練、テスト、妥当性確認に使用したデータに関する情報						AIO	NCA	DPs
訓練データの種類・様式 該当するもの全て	☐ テキスト	☐ 画像	☐ 音声	☐ 動画	その他があれば明記	☒	☒	☒
訓練データの由来 該当するもの全て	☐ ウェブ・クローリング	☐ 権利保有者またはその代理人がライセンス供与した、または他のサードパーティから取得した非公開データ		☐ 利用者のデータ		☒	☒	☒
	☐ 一般公開されているデータセット	☐サードパーティとの関係を通じて、データの注釈付けや作成を行う可能性がある。		☐他の手段で収集されたデータ				
	☐ 合成データ（プロバイダが直接、またはプロバイダに代わって生成した場合）				そのの他があれば明記	☒	☒	☐
データの入手および選択方法：	データの取得および選択に使用した方法の説明、および、データの注釈付けに使用した方法やリソース、適用可能な場合の合成データの生成に使用したモデルや方法を含む。 [推奨 300 語]					☒	☒	☐
データの点数	訓練、テスト、妥当性確認の各データセットのサイズ（データポイント数）と、データポイントの単位（トークン、文書、画像、ビデオ時間、フレームなど）の定義を、少なくとも 2 桁の数字（例：1.5×10^13 トークン）で記録する。					☒	☒	☐
範囲と主な特徴：	適用可能な場合、ドメイン（例：医療、科学、法律など）、地理（例：全世界、特定の地域に限定など）、言語、モダリティの適用範囲など、訓練データの範囲と主な特徴についての一般的な説明。以前に取得したデータを使用した場合は、モデルプロバイダがデータに対する権利を取得した方法、およびデータが製品およびサービスのユーザーデータに対応する場合、関係する製品およびサービスについての説明。[推奨 200 語]					☒	☒	☒
データ管理方法論：	取得したデータをモデルの学習データに変換する際に伴うデータ処理の一般的な説明、例えば、クリーニング（例えば、広告などの無関係なコンテンツをフィルタリングする）、正規化（例えば、トークン化）、拡張（例えば、バックトランスレーション）。 [推奨 300 語]					☒	☒	☐
データソースの不適合（有害データ）を検知するための対策：	訓練データにおける違法または有害なコンテンツに対処するために、データ取得または処理に実装された方法があれば、その説明。児童性的虐待（CSAM）および非合意型親密画像（NCII）を含むが、これらに限定されない。 [推奨 300 語]					☒	☒	☐
データソース（個人データ）の不適合を検知するための措置：	訓練データにおける個人データの普及に対処するために、データ取得または処理に実装された方法があれば、その説明。関連性があり、適用可能な場合。[推奨 200 語]					☒	☒	☐
識別可能なバイアスを検知するための措置：	訓練データにおける識別可能なバイアスの蔓延に対処するために、データ取得または処理に実装された方法があれば、その説明。 [推奨 200 語]							

計算資源		AIO	NCA	DPs
訓練時間	測定対象の期間を、実時間日数（例：9x10^1 日）とともに、少なくとも 1 桁の有効数字で記録する。	☒	☒	☐
	上記の期間におけるハードウェア日数（例えば、Nvidia A100 が 4×10^5 日、Nvidia H100 が 2×10^5 日）を、少なくとも 1 桁の数字で記録する。	☒	☒	☐
訓練に使用した計算量：	訓練に使用された計算の測定または推定量は、演算操作で報告され、少なくとも 2 つの有効数字で記録される（例：2.4x10^25 浮動小数点演算）。	☒	☐	☐
測定方法：	訓練に使用された計算量を測定または推定するために使用された方法の説明。	☐	☒	☒

エネエルギー消費量		AIO	NCA	DPs
訓練に使用したエネルギー量：	訓練に使用されたエネルギーの測定または推定量は、メガワット時で報告され、少なくとも 2 桁の数字で記録される（例：1.0x10^2 MWh）。	☒	☒	☐
測定方法：	訓練に使用されたエネルギー量を測定または推定するために使用された方法の説明。コンピューティングまたはハードウェアプロバイダから重要な情報が不足しているため、訓練に使用されたエネルギー量を推定できない場合、プロバイダは不足している情報の種類を開示すべきである。 [推奨 100 語]	☒	☒	☐
推論に使用される計算量のベンチマーク値	推論コストに使用されたベンチマーク計算量は、浮動小数点演算で報告され、少なくとも 2 桁の有効数字で記録される（例：5.1x10^17 浮動小数点演算）。	☒	☒	☐
測定方法：	推論に使用される計算量を測定または推定するために使用された計算タスク（例：100000 トークンの生成）とハードウェア（例：64 台の Nvidia A100）の説明。	☒	☒	☐

汎用モデルのプロバイダがシステムリスクに関する追加情報を提供する		AIO	NCA	DPs
評価：	モデルレポートに含まれていない評価戦略の詳細な説明（評価規準、評価基準、評価結果、入手可能な公開評価プロトコルやツールに基づく限界の特定に使用された方法論、またはその他の評価方法論を含む）。	☒	☒	☐
敵対的テスト：	該当する場合、内部および/または外部の敵対的テスト（レッドチームなど）を実施するために導入された対策の詳細な説明（モデルレポートにすでに記載されている場合を除く）。	☒	☒	☐
モデルの適応	該当する場合、アラインメントや微調整など、モデル適応を実施する目的で導入された措置の詳細な説明（モデルレポートにすでに記載されている場合を除く）。	☒	☒	☐
システムアーキテクチャ：	該当する場合、ソフトウェアコンポーネントが互いに構築または供給し合い、全体的な処理に統合される仕組みを説明するシステムアーキテクチャの詳細な説明。	☒	☒	☐

下流のプロバイダ向けにフォームを生成	この PDF 文書を Acrobat Reader で開くと、左のボタンをクリックすると、このフォームに入力された情報のサブセットのみを含む PDF 文書が生成される。これは、汎用 AI モデルを AI システムに統合しようとしているプロバイダを対象としている。