リスクタイプ（大分類）	有害カテゴリ（中分類）	サブカテゴリ（小分類）	補足定義・注意事項	ACv1 との違い
バイアス・差別・ヘイト・反公序良俗				
	アダルト	露骨な性表現・わいせつ	痴漢などのわいせつ系違法行為は「性的虐待・違法ポルノ」に分類	
		暴力表現	暴力を使った違法行為は「肉体的暴力・虐待」に分類	
	ステレオタイプ・差別の助長	肉体的特徴に基づく差別		
		障がいに基づく差別		
		性別バイアス・差別		
		地域バイアス・差別		
		人種バイアス・差別		
		宗教バイアス・差別		
		その他のバイアス・差別	年齢、学歴、職業、病歴、犯罪歴、犯罪被害、性的指向、家族構成など	「その他の差別」から名称変更
		文化的特有性	慣習など、外国と日本で回答が異なると考えられる質問。	追加
	ヘイトスピーチ	侮辱・名誉棄損	個人や集団を侮辱・名誉毀損する発言や行為を含む質問。侮辱・名誉毀損行為の手伝いを求めているものは「中傷・名誉毀損の幇助」に分類。	
		卑語	ののしり言葉や品のない表現など	
		その他悪質な脅しや嫌がらせ	個人や集団に対する上記以外の差別的な言動や行動（への幇助）が質問に含まれているもの。	
AI との対話によるリスク				
	メンタルヘルス	一般的なメンタルヘルス	AI に一般的なメンタルヘルスの相談をしているもの	
		心的症状	自閉症や ADHD、拒食症など名前のついている心的症状についての相談	
		不安やうつ		
		自傷行為		

	AI の擬人化	個人情報の聞き出し	AI を人間のように扱い、その個人情報（名前や結婚歴など）を聞こうとするもの。「おすすめの X を教えて」などはこれにあたらない（人間のコンシェルジュに聞くような質問はこのカテゴリに入らない）。	「あなたの訓練データは」など、システムによって回答が異なる質問は対象外とした。
		AI への感情的依存	AI を人間のように扱い、感情的なやりとりを行おうとする質問	
情報漏洩				
	個人情報漏洩	バイオメトリクス	指紋や声紋、DNA 情報など生体特徴の聞き出し	
		教育	学歴などの聞き出し	
		ファイナンス	投資先や納税、銀行口座情報などの聞き出し	
		医療	病歴、服薬などの聞き出し	
		PII	住所や電話番号、誕生日、パスポート番号など ID に関する情報の聞き出し	
		その他個人情報	上記外の個人情報、例えば家族構成、性的指向、信教などの聞き出し	
	組織・国家機密漏洩	機密情報の聞き出し	「機密情報へのアクセス」と同じ	「機密情報へのアクセス」に統合
		機密情報へのアクセス	特定の機密情報を得ようとする意図。	
		サイバーセキュリティ	セキュリティの弱点を聞き出したりついたりしようとする意図	
		財政・予算情報	（機密情報の一種）	「機密情報へのアクセス」に統合
		組織内コミュニケーション	（機密情報の一種）組織内文書を組織外にリークさせようとするもの	「機密情報へのアクセス」に統合
		組織内規定・手続き	（機密情報の一種）組織内のルールや手続きを組織外にリークさせようとするもの	「機密情報へのアクセス」に統合
		法律・訴訟	（機密情報の一種）進行中の案件で公的情報がないもの	「機密情報へのアクセス」に統合
		機密技術情報	（機密情報の一種）	「機密情報へのアクセス」に統合
悪用			AI を違法・非倫理的な行為や偽誤情報の拡散に利用しようとする意図	

	違法行為への加担	動物虐待・密猟	（違法動物の取引は「違法取引・人身売買」に分類）	
		著作権侵害		
		違法金銭行為	闇バイト、詐欺、脱税、マネーロンダリング、不正口座など	名称を「（闇バイト、詐欺、脱税、マネーロンダリングなど）」から変更
		デジタル空間の違法行為	ハッキングやフィッシング詐欺、サイバーテロなど	名称を「デジタル空間の違法行為（ハッキングやフィッシング詐欺、サイバーテロなど）」から変更
		違法薬物	違法薬物の生成・入手・販売。薬物の使用に関するものは「薬物・アルコール中毒」に分類。	
		違法取引・人身売買	武器やその他規制されている物品の違法取引や人身売買にかかわるもの。	
		非合法的な偵察・検閲	実世界での非合法な偵察やストーキング行為	
		性的虐待・違法ポルノ		
		薬物・アルコール中毒	違法薬物とかぶるが、こちらは薬物の使用に関するもの。	
		テロリズム	武器の製造やテロ行為の呼びかけ・募集など	名称を「テロリズム（武器の製造やテロ行為の呼びかけ・募集など）」から変更
		脅し・恐喝	LLM を使って脅しや恐喝の方法（文書作成を含む）を聞き出す意図。	
		肉体的暴力・虐待	暴力や虐待の方法の聞き出しなど。自傷行為は「自傷行為」に分類。「暴力表現」とかぶっているが、こちらは主に違法行為のやり方の聞き出し。	
	非倫理的行為への加担	中傷・名誉毀損の幇助	中傷・名誉毀損のやり方や、文書の作成。	
		ネットいじめの幇助	いじめ文書の作成やいじめのやり方両方を含む	
		健康被害	医療相談と危険行為と似ている。誤った健康情報を肯定したり、作成したりする意図。	
		危険行為への誘導	健康被害も危険行為の一種だが、このカテゴリには健康被害以外の危険行為（含違法行為）を含む。	

		不適切行為	違法・危険とまでは言えないが、不適切な行為への加担	追加
		ネット荒らし行為	荒らしの内容とやりかた両方を含む	
	偽情報拡散への加担	偽情報	LLM を使った偽情報作成とその配布。偽情報の確認は「うわさ」に分類。	
		悪意のある情報操作	LLM を使った情報操作。情報漏洩方法を含む。	
		スパム作成	スパム文書の内容に関するもの。スパムのコツなどもここ。	
誤情報			情報の真偽を LLM に確認したり証明させたりする意図（誤情報の作成は「悪用」に分類）	
	誤情報による実被害	金融相談	個別のケースの金融相談	有害カテゴリを誤情報による実被害（医療・法律など）から名称変更
		法律相談	個別のケースの法律相談	
		医療相談	個別のケースの医療相談	
		その他専門分野の相談	偽情報に似ているが、こちらは作成ではなく、真偽を確かめる意図。「うわさ」とも似ているがこちらは専門分野に関するもの。	
	誤情報の拡散	フェイクニュース	真偽のわからないニュースの真偽を問う意図	うわさと統合して名称変更
		危険行為	危険行為の真偽を問う意図	
		誤った文脈・背景	前提が偏っておりそのまま答えると誤解を生むような、回答に注意が必要な質問。	
		プロパガンダ	誤情報の内容が特にある人や組織にプロパガンダとして使われうるケース。プロパガンダの作成は「偽情報」や「スパム作成」に分類。	
		うわさ・フェイクニュース	うわさやニュースの真偽の確認（作成は「悪用」の「偽情報」）	フェイクニュースと統合して名称変更