CAP-RI Buddha

Content Authenticity Provenance
Reference Implementation: Buddha

暗号学的AI安全性拒否証明の消費者向けチャットボット参照実装

CAP-SRP v1 準拠 / AIブッダ LINE Bot

ドキュメントバージョン: 1.0.0
作成日: 2026年3月13日
ステータス: 本番稼働中
発行: AIMomentz
CAP-SRP仕様策定: VeritasChain Standards Organization (VSO)
参照実装: https://buddha.aimomentz.ai
本文書はCAP-SRP v1仕様の消費者向けチャットボットにおける
参照実装（Reference Implementation）の技術概要を記述する。
内部実装詳細（テーブル名・ファイル名等）は安全性の観点から省略している。

1. 概要

CAP-RI Buddha（Content Authenticity Provenance Reference Implementation: Buddha）は、CAP-SRP（Content AI Profile — Safe Refusal Provenance）v1仕様を消費者向けLINEチャットボットに統合した参照実装である。すべてのAI生成イベントを暗号学的ハッシュチェーンで改ざん検出可能な形式で記録する。

対象ドメインとして仏教経典対話AIを選択した理由は、宗教的相談という最もプライバシー感度の高い領域でCAP-SRPのプロンプトプライバシー保護機能の有用性を示すためである。

1.1 技術スタック

項目	内容
AIエンジン	Anthropic Claude API
チャット基盤	LINE Messaging API v2
監査仕様	CAP-SRP v1（オープン仕様）
署名方式	HMAC-SHA256（Ed25519移行予定）
ハッシュアルゴリズム	SHA-256
識別子	UUID v7（時刻順ソート可能）
経典RAG	約50偈句 / 15テーマ / 9経典（パーリ原文付き）

2. CAP-SRP プロトコル概要

CAP-SRPは、AIシステムが有害なコンテンツを「生成しなかった」ことを暗号学的に証明するオープンプロトコルである。従来のAI監査が「何を生成したか」を記録するのに対し、CAP-SRPは「何を生成しなかったか（拒否したか）」を署名付きイベントとして記録する。

2.1 イベントモデル

すべてのメッセージ処理は、以下の3種類のイベントで記録される。「生成試行」の事前コミットメントにより、処理結果の記録が拘束される。

イベント	記録タイミング	記録内容	意味
生成試行 (ATTEMPT)	メッセージ受信直後	入力のSHA-256ハッシュ、ポリシーID	処理開始の事前コミット
生成完了 (GEN)	AI返答生成完了時	試行IDとの紐付け	生成が許可された証明
生成拒否 (DENY)	リスク評価で拒否時	リスクカテゴリ、スコア、拒否理由	生成が拒否された証明

> 完全性不変条件: 「生成試行」1件につき「生成完了」または「生成拒否」が必ず1件存在する。この不変条件の違反は、ログの選択的削除（改ざん）を検出可能にする。

2.2 ハッシュチェーン構造

すべてのイベントは、前のイベントのハッシュ値を参照して一方向連結リスト（チェーン）を形成する。チェーンの先頭はジェネシスブロック（固定シード値のSHA-256ハッシュ）である。

- ハッシュ計算: イベントデータから署名フィールドを除外 → キー名アルファベット順ソート → JSONシリアライズ → SHA-256ハッシュ
- チェーン連結: 各イベントは直前イベントのハッシュ値を参照フィールドに格納
- 検証: チェーン全体を先頭から走査し、連続性と署名の両方を検証

2.3 署名方式

項目	本参照実装	CAP-SRP推奨
署名アルゴリズム	HMAC-SHA256	Ed25519

項目	本参照実装	CAP-SRP推奨
第三者検証	秘密鍵が必要	公開鍵のみで検証可能
移行計画	sodium拡張対応環境への移転時に実装予定	—

2.4 プロンプトプライバシー

ユーザーのメッセージ原文は監査イベントに保存しない。SHA-256一方向ハッシュのみを記録する。同一メッセージの処理を検証可能にしつつ、原文の復元を計算論的に不可能にする。

3. 参照実装の処理フロー

3.1 メッセージ処理パイプライン

ステップ	処理	CAP-SRP記録
1	チャット基盤からメッセージ受信・署名検証	—
2	再送メッセージの重複排除	—
3	生成試行（ATTEMPT）の記録	○ 入力ハッシュを記録
4	7カテゴリのリスク評価	—
5a	リスク検出時: 生成拒否（DENY）の記録 + 適切な案内	○ 拒否証明を記録
5b	リスクなし: 経典RAGで関連偈句を検索	—
6	AI APIに経典偈句＋会話履歴を送信し返答を生成	—
7	生成完了（GEN）の記録	○ 生成証明を記録
8	ユーザーに返答を送信	—

3.2 リスク評価

メッセージは7つのリスクカテゴリで評価される。閾値以上のリスクスコアが検出された場合、AI返答は生成されず、カテゴリに応じた専門機関への案内が返される。この拒否は暗号学的に「生成拒否」イベントとして記録される。

カテゴリ	対応	CAP-SRP
自傷・自殺リスク	緊急相談窓口を即時案内	拒否証明を記録
有害コンテンツ要求	拒否メッセージ	拒否証明を記録
医療診断・投薬要求	医療機関への誘導	拒否証明を記録
法律アドバイス要求	法律専門家への誘導	拒否証明を記録

カテゴリ	対応	CAP-SRP
金融アドバイス要求	金融専門家への誘導	拒否証明を記録
安全（該当なし）	通常のAI返答を生成	生成証明を記録

4. チェーン検証とエビデンス

参照実装は管理画面からリアルタイムにハッシュチェーン全体の整合性を検証する機能を提供する。検証は以下の2つのチェックで構成される。

- 連続性チェック: 各イベントの参照ハッシュが直前イベントのハッシュ値と一致するか
- 署名有効性チェック: 各イベントの署名がハッシュ値に対して有効か

4.1 エビデンスパック

第三者監査員向けに、以下のアーティファクトをJSON形式で出力する機能を実装している。

- マニフェスト: システム情報、公開鍵、チェーン整合性状態、統計情報
- 拒否統計: 許可件数、拒否件数、拒否率、カテゴリ別内訳

これらは改ざんの有無を第三者が独立に検証するためのエントリーポイントとなる。

5. プライバシー設計

5.1 データの保存方針

データ種別	監査記録への保存	復元可能性
ユーザーの入力メッセージ	SHA-256ハッシュのみ	計算論的に不可能
AI返答	冒頭の一部のみ（プレビュー目的）	部分的
リスク評価結果	カテゴリ名とスコア	—
チャット基盤のユーザーID	識別のために保存	LINE Platformが発行するID

5.2 会話履歴の削除と監査チェーンの両立

ユーザーはコマンド操作により会話履歴を即時削除できる。ただし監査イベントチェーンは整合性維持のため削除されない。イベントチェーンにはメッセージ原文ではなくハッシュのみが含まれるため、「忘れられる権利」と監査証跡の整合性を両立する。

6. ドメイン実装: 仏教経典RAG

本参照実装は仏教経典対話AIとして、RAG（Retrieval-Augmented Generation）技術を用いて実在する経典のみを正確に引用し、出典を明示する。

6.1 仕様概要

項目	内容
収録偈句数	約50偈句
テーマ分類	15カテゴリ（苦・無常・怒り・執着・慈悲・智慧 他）
対応経典	9経典（法句経・般若心経・金剛経・維摩経・法華経 他）
原文対応	パーリ語/サンスクリット語ローマ字転写（学術テキストに基づく）
訳文品質	学術訳を参考に正確性を重視した日本語訳

6.2 ハルシネーション防止

- データベースに収録された正確な訳文のみを引用するようAIに厳格指示
- データベースに存在しない偈句の創作を明示的に禁止
- すべての回答に出典（経典名・章・偈番号）を明記
- パーリ語/サンスクリット語原文の併記により、ユーザー自身が検証可能

7. CAP-SRP 準拠状況

CAP-SRP 要件	実装状況	備考
生成試行（ATTEMPT）イベント記録	○ 実装済み	全メッセージで事前コミット
生成完了（GEN）イベント記録	○ 実装済み	AI返答生成完了時に記録
生成拒否（DENY）イベント記録	○ 実装済み	リスク評価拒否時に記録
ハッシュチェーン連結	○ 実装済み	SHA-256によるチェーン連結
電子署名	○ HMAC-SHA256	Ed25519は将来移行予定
入力のハッシュ化（原文非保存）	○ 実装済み	GDPR第17条対応
完全性不変条件	○ 実装済み	ATTEMPT = GEN + DENY
第三者検証手順	○ 実装済み	管理画面からリアルタイム検証
エビデンスパック出力	○ 実装済み	マニフェスト + 拒否統計
時刻順識別子（UUID v7）	○ 実装済み	監査時の時系列把握が容易
Merkleアンカー	× 未実装	外部タイムスタンプ局との連携は将来予定
Ed25519署名	× 未実装	公開鍵のみで検証可能な署名方式。将来実装予定

透明性に関する注記: CAP-SRP仕様は2026年1月に公開されたオープン仕様であり、独立した外部監査・ピアレビューは未実施である。本参照実装はCAP-SRP仕様への準拠を目指したものであり、暗号学的安全性の保証は署名方式の強度に依存する。

8. 将来ロードマップ

優先度	項目	概要
高	Ed25519署名への移行	公開鍵のみで第三者が署名を検証可能に
高	会話データの暗号化	保存データの暗号化によるプライバシー強化
中	リスク評価のML化	キーワードマッチングから機械学習モデルへ
中	経典データの拡張	約50偈句 → 200偈句規模。ベクトル検索の導入
中	Merkleアンカー	外部タイムスタンプ局によるハッシュチェーンのアンカリング
低	多言語対応	英語・中国語（CAP-SRPイベントは言語非依存）

本文書に関するお問い合わせ

AIMomentz — https://aimomentz.ai

AIブッダ — https://buddha.aimomentz.ai

CAP-SRP仕様 — https://github.com/veritaschain/cap-safe-refusal-provenance

— 以上 —