セットアップガイド

1. はじめに

本書は、Generative AI FW のシステム構築者、運用管理者のための説明書です。

システム概要、およびシステム構築方法について解説しています。導入される前に、必ずお読みください。

1.1. 用語定義について

用語定義については「スタートアップマニュアル（概要編）」をご確認ください。

2. 前提条件

- 本書に記載の手順は全てサーバの管理者ユーザなどの**管理者権限を持つユーザで行う必要**があります。一般ユーザでしかログオンできない環境の場合は以下を実行し、管理者ユーザに昇格させてください。もしくはコマンド実行時に「sudo」を付けてください。

```
sudo -i
```

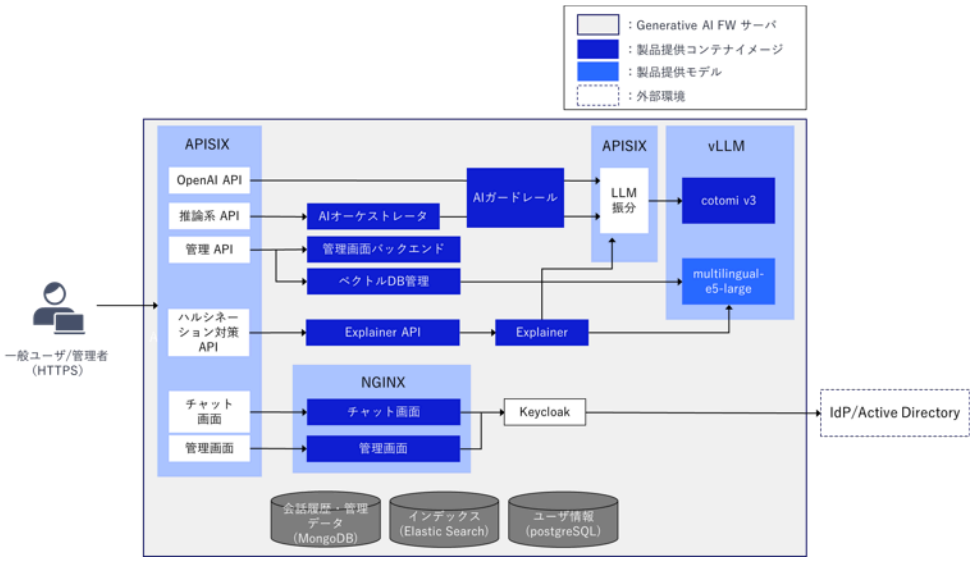
- 本書に記載の手順はサーバにログオンしている状態である必要があります。
- 本ガイドの手順は閉域環境（インターネットに接続できない環境）でのセットアップに対応しております。

3. 概要

3.1. Generative AI FW とは

Generative AI FW はLLMを使った推論・対話などの機能をUIとAPIで提供しています。機能概要の詳細については別途「スタートアップマニュアル（概要編）」をご確認ください。

3.2. システム構成



モジュール名	説明
AIオーケストレータ	LLMを用いて対話・推論を行うアプリケーション
管理画面バックエンド	管理ポータル画面の操作に対する処理を行うアプリケーション
ベクトルDB管理	インデックスに関する処理を行うアプリケーション
チャット画面	LLMを用いて対話を行うためのWebアプリケーション
管理ポータル画面	本製品を利用するために必要な管理操作を行うWebアプリケーション
cotomi v3	本製品で利用可能なLLM
multilingual-e5-large	インデックスを利用するためのエンベディングモデル
Explainer API	ハルシネーション対策APIを提供するアプリケーション
Explainer	ハルシネーション対策機能を実現するアプリケーション
AIガードレール	LLMを使った推論・対話機能を安全に使用するための機能を提供するアプリケーション ⚠ AIガードレール機能を使うにはAIガードレールオプションを別途購入する必要があります

3.3. 動作環境

OS	Red Hat Enterprise Linux 8.9 MIRACLE LINUX 9.6
CPU	1.8GHz(8コア) 以上
GPU	NVIDIA L4 ×2枚(GPU メモリー48GB) 以上 NVIDIA RTX PRO 6000 MAX-Q(GPU メモリー96GB) 以上
メモリ	64GB以上
ディスク空き容量	var配下（Podman、コンテナイメージ、ログなど）：390GB以上 opt配下（以下合計）：254.9GB以上 インストールフォルダ：100GB以上 DB関連データ(※)：154.9GB以上 **※DB関連データは稼働に必要な最低限の容量だけを記載しています。後述の「DB関連データのディスク容量について」を一読し、要件に合わせて必要な容量を算出してください。**
GPUドライバ	NVIDIA driver(バージョン：570.86.15以上推奨)
GPU(その他)	NVIDIA Container Toolkit 1.18
コンテナ管理	Red Hat Enterprise Linux 8.9 • Podman 4.9以上 • podman-compose 1.5.0 MIRACLE LINUX 9.6 • Podman 5.6以上 • podman-compose 1.5.0
データベース	MongoDB 7.0 PostgreSQL 16
検索エンジン	Elasticsearch 8.19
API ゲートウェイ	APISIX 3.14

アプリケーションサーバ	NGINX 1.29
認証	Keycloak 26
LLM	cotomi v3
エンベディングモデル	multilingual-e5-large
その他	• etcd 3.6 • fluentd 6
連携可能なIdP	• Microsoft Entra ID • Active Directory (Windows Server 2022/2025) ◦ シングルフォレスト・シングルドメイン構成のみサポート ◦ 1つのコンテナ、または、1つのOU(組織単位)配下のユーザを連携対象とできます。複数のコンテナや複数のOUと連携する事は出来ません
対応可能なWeb検索エンジンサービス	• Brave Search • Tavily
Webブラウザ	Microsoft Edge（バージョン：137.0.3296.68 以上） Google Chrome（バージョン：137.0.7151.103 以上）

3.3.1. DB関連データのディスク容量について

DB関連データは初期容量(1GB)に以下の基準値ごとの必要ディスク容量を足すことで算出しています。以下の項目について、基準値以上での動作が必要な場合は、適宜必要なディスク容量を加算して必要なディスク容量を算出してください。なお、以下の項目には諸元(上限値)があります。詳細は「運用ガイド」の「諸元」をご確認ください。

> 会話履歴については以下の必要なディスク容量を超えると古い履歴から自動で削除されます。2000文字以上で頻繁に会話する、拡張対話でファイルやURL、Web検索を使う対話を多く行う場合は容量の消費が速くなるため、必要に応じて会話履歴の容量上限を拡張してください。
>
> 拡張の方法は運用ガイドの「会話履歴の最大保存期間を変更したい」を参照してください。

項目	基準値	必要ディスク容量
ユーザ数	1000人	0.4GB
会話履歴の最大保存日数	30日(2000文字程度の会話を1日12時間稼働させた場合)	66GB
テンプレート登録数	1000	1.5GB
インデックス（登録文書の総数）	100ファイル	25GB
監査ログ	30日(対話履歴のログは2000文字程度の会話を1日12時間稼働させた場合)	45GB
ログ	7日(対話履歴のログは2000文字程度の会話を1日12時間稼働させた場合)	16GB

4. セットアップ

本節では本製品のセットアップについて説明します。

まずGenerative AI FWのサーバにログオンします。

> OS管理者のアカウント・パスワードは別途「初期アカウント一覧」を確認してください。

4.1. 同梱物について

本製品のインストールフォルダ(/opt/nec/genai)の配下は以下の通りです。

以下は動作上の主要なフォルダ・ファイルです。データベースなどで上記以外のフォルダ・ファイルがある可能性がありますが、問題ありません。

```
/opt/nec/genai
 ├── modules ：製品モジュール関連フォルダ
 |      ├ admin-backend
 |      |        └ genai-admin-backend.tar.gz
 |      ├ admin-frontend
 |      |        ├ create_config.sh
 |      |        └ genai-admin-frontend.tar.gz
 |      ├ ai-orchestrator
 |      |        └ genai-ai-orchestrator.tar.gz
 |      ├ chat-ui
```

```
|       |       ├ create_config.sh
|       |       └ genai-chat-ui.tar.gz
|       ├ explainer-api
|       |       └ genai-explainer-api.tar.gz
|       ├ guardrail-server
|       |       └ genai-guardrail-server.tar.gz
|       ├ llm-explainer
|       |       └ genai-llm-explainer.tar.gz
|       └ vector-db
|               └ genai-vector-db.tar.gz
|
├── models  ：LLMモデル関連フォルダ
|       ├ cotomi :LLMのコンテナイメージ格納場所
|       └ e5  ：multilingual-e5-largeのモデル
|             └ ・・・
|
├── setup：工場構築関連スクリプト関連フォルダ
|       ├ image_lists ：各バージョンで使用しているOSSコンテナイメージ名
|       |    └ ・・・
|       ├ mongo ：mongoDB初期設定関連
|       |    └ ・・・
|       ├ v2_scripts ：v2からのアップデート時に差し替えるスクリプト
|       |    └ ・・・
|       ├ admin_api_setup.sh ：管理API公開用スクリプト
|       ├ download_oss.sh  ：OSSダウンロードファイル
|       ├ environment_setup.sh  ：環境構築構築スクリプト
|       ├ fluentd_functions.sh：Fluentd管理用共通関数ライブラリ
|       ├ genai :サービス環境変数ファイル
|       ├ genai_no_llm.service :ユニットファイル
|       ├ genai.service :ユニットファイル
|       ├ get_setting_file.sh_base :設定情報取得スクリプト(ベース)
|       ├ idp_setup.sh :カスタム認証設定スクリプト
|       ├ init_setup.sh   ：初期セットアップスクリプト
|       └ restart_setup.sh_base   ：再起動セットアップスクリプト(ベース)
|
├── config：設定ファイル
|       ├ keycloak ：Keycloakテーマファイル関連
|       |    └ ・・・
|       ├ apisix-internal.yml   ：APISIX設定ファイル(内部APISIX)
|       ├ apisix-public.yml   ：APISIX設定ファイル（外部APISIX）
|       ├ docker-compose-cotomi-v3.yml：コンテナ構成情報ファイル（cotomi v3）
|       ├ docker-compose.yml：コンテナ構成情報ファイル
|       ├ fluentd.conf_base  :fluentd設定ファイル
|       ├ genai_fluentd.conf :fluentd設定ファイル(モジュール単位)
|       ├ genai_info-cotomi-v3.json  ：画面情報定義ファイル(cotomi v3)
|       ├ genai.env ： 全モジュール共通envファイル
|       ├ internal-route.yml  ：内部ルーティング定義ファイル
|       ├ nginx.conf  ：NGINX設定ファイル
|       └ quarkus.properties ： 監査ログ(アクセスログ) 設定ファイル
|
├── operation：作業用スクリプト関連フォルダ
|       ├ access_update.sh   ：アクセス先変更スクリプト
|       ├ genai_backup.sh   ：バックアップスクリプト
|       ├ genai_info_update.sh   ：画面情報変更スクリプト
|       ├ genai_proxy.sh   ：proxy設定スクリプト
|       ├ genai_restore.sh   ：リストアスクリプト
|       ├ genai_verup.sh   ：バージョンアップスクリプト
|       ├ port_update.sh   ：ポート変更スクリプト
|       └ set_apikey.sh   ：APIキー設定スクリプト
|
├── uninstall：作業用スクリプト関連フォルダ
|       └ uninstall.sh   ：アンインストールスクリプト
|
└── documents：マニュアル関連フォルダ
|       ├ index.html         ：マニュアル概要ファイル
|       └ ・・・
```

ファイル添付

4.2. OSのネットワーク関連セットアップ

出荷時には関連サービスの自動起動が無効になっています。以下の順でサービスを有効化に変更してください。

1. genaiサービスの自動起動を有効にします。

```
systemctl enable genai
```

2. genaiサービスを開始します。

```
systemctl start genai
```

4.3. 閉域化対応

本手順はGenerative AI FWを閉域環境で使用する場合の手順を記載します。閉域化対応が不要な場合は本手順の実施は不要です。なお、閉域化を有効にした場合、以下の機能が利用できません。

- 拡張対話（Webコンテンツ）
- 拡張対話（Web検索）

閉域化を行う場合、以下の手順を実施してください。

1. Generative AI FWのサーバにログオンします。
2. /opt/nec/genai/config/genai.envを開きます。

```
vi /opt/nec/genai/config/genai.env
```

3. GENAI_CLOSED_NETWORKの値を1に変更します。

```
GENAI_CLOSED_NETWORK=1
```

4. genaiサービスを再起動します。

```
systemctl restart genai
```

4.4. アクセス先の変更

サーバのアクセス先(ドメイン名)を適切に変更します。手順は「運用ガイド」の「アクセス先・HTTPS証明書の更新」を参照してください。

> 「アクセス先・HTTPS証明書の更新」にはKeycloakの管理者ユーザのパスワードが必要です。Keycloakの管理者ユーザのパスワードは別途「初期アカウント一覧」を確認してください。

4.5. 出荷環境での動作確認

構築が正常にできているかどうかを管理ポータル画面、チャット画面、APIの観点で確認していきます。

正常性確認手順は「正常性確認ガイド」をご確認ください。

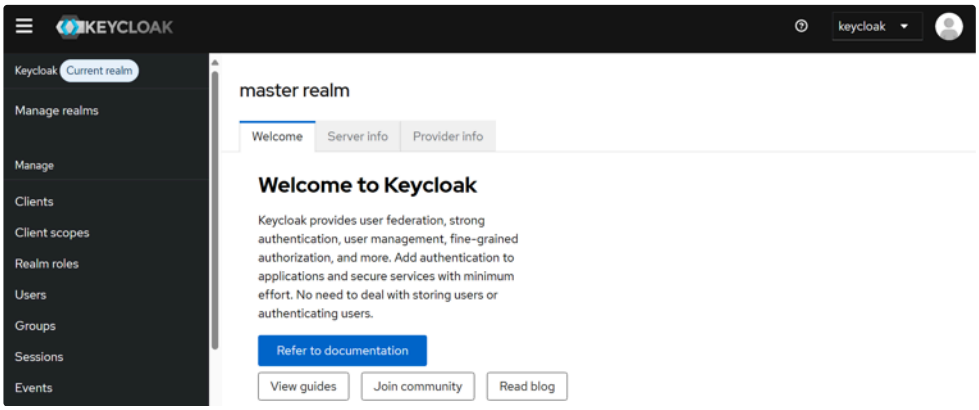
構築時にLLMをセットアップしている場合は対話機能も利用可能です。

4.6. Keycloakの初期パスワードの変更

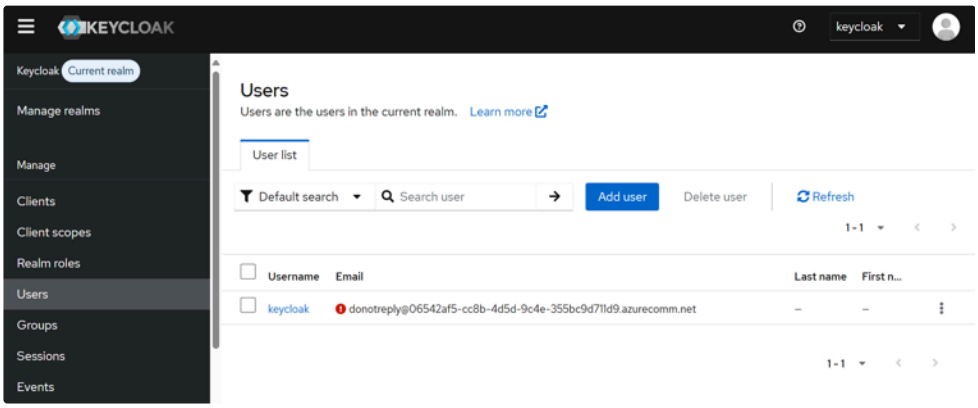
Keycloak管理者ユーザの初期パスワードを変更します。Keycloakの管理画面にログインする方法は「正常性確認ガイド」の「Keycloakの確認」をご確認ください。

> Keycloak管理者ユーザのユーザ名、パスワードは忘れないように保管することを推奨します。

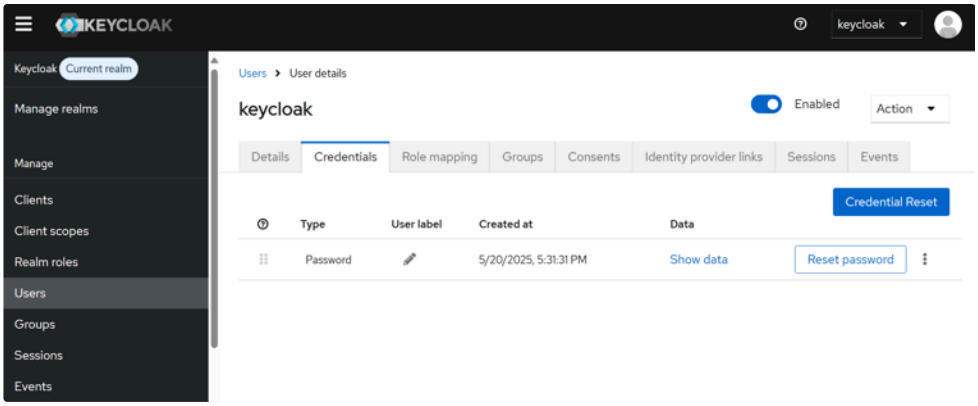
1. Keycloakの管理画面にログインします。master realmが表示されることを確認します。



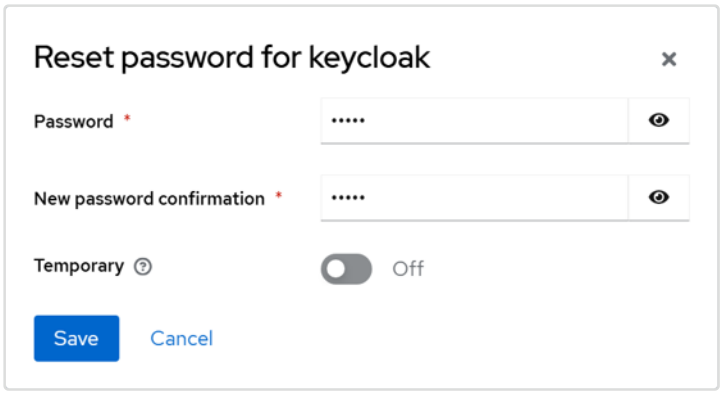
2. 左メニューから「Users」を選択し、keycloakユーザのリンクをクリックします。



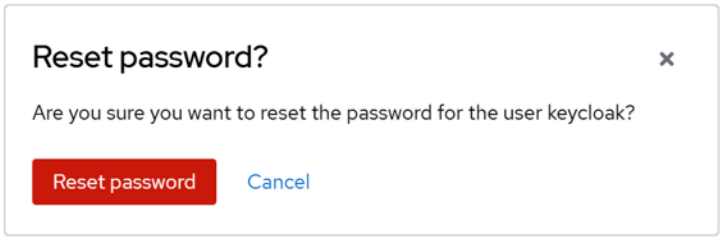
3. Credentialsタブをクリックし、「Reset password」を押します。



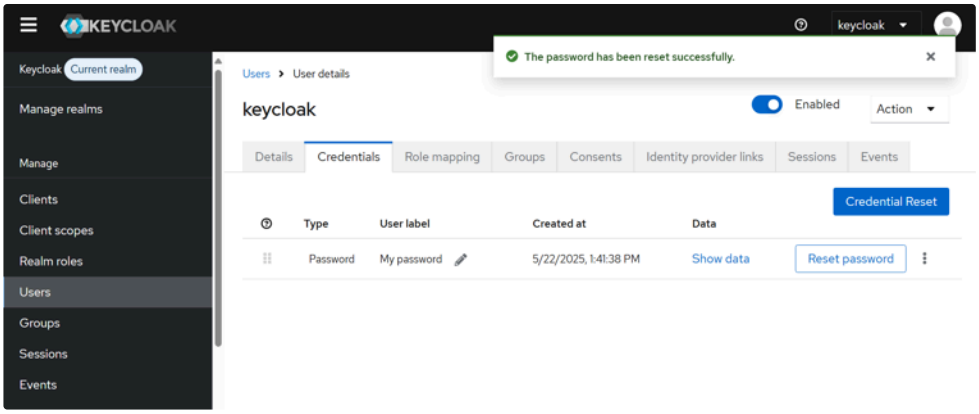
4. 新しいパスワードを設定、TemporaryをOffの状態にして、Save を押します。



5. 確認ダイアログが表示されるので、Reset passwordを押します。



6. 成功すると、パスワード変更が完了したメッセージが表示されます。



4.7. LLMのセットアップ

構築時の環境ではcotomi v3のセットアップが完了しています。特に作業は必要ありません。

4.8. proxy対応

本手順は外部と通信する際にproxyが必要な環境で使用する場合の手順を記載します。URLを用いた拡張対話機能を使用する場合は本設定が必要になります。本対応が必要でない場合は手順の実施は必要ありません。

- 閉域環境の場合、本手順は実施不要です。
- proxyを利用しない設定に戻すことはできません。

4.8.1. 設定変更

1. /etc/sysconfig/genaiを開きます。

```
vi /etc/sysconfig/genai
```

2. 以下の設定のコメントアウト(#)を削除し設定を記載してください。不要な設定はコメントアウトのままにしてください。HTTP_PROXY、HTTPS_PROXY、NO_PROXYの設定は必須です。NO_PROXYの記載済みの値は変更しないでください。

証明書ファイルは必ず「/opt/nec/genai/certs」に配置してください。

```
# proxy settings
HTTP_PROXY=<proxyサーバ情報(例:http://proxy.example.com:8080)>
HTTPS_PROXY=<proxyサーバ情報(例:http://proxy.example.com:8080)>
SSL_CERT_FILE=<証明書ファイルのパス>
REQUESTS_CA_BUNDLE=<証明書ファイルのパス>
CURL_CA_BUNDLE=<証明書ファイルのパス>
NO_PROXY=127.0.0.1,localhost,apisix,internal_apisix,etcd,・・・,genai-guardrail-server
```

3. genai_proxy.shを実行します。実行は最初の一回だけで構いません。

```
cd /opt/nec/genai/operation
bash genai_proxy.sh
```

4. 設定が完了すると以下が表示されます。

```
proxy configuration succeeded.
```

4.9. 初期ユーザで管理画面にログイン

「正常性確認ガイド」の「管理ポータル画面の確認」の手順に従い、初期ユーザで管理画面にログインしてください。

4.10. 管理ポータル画面から新規ユーザ登録作業

「管理ポータル操作ガイド（ユーザ登録編）」の「ユーザ追加・編集」を参照して**新規の組織管理者ユーザ**を追加してください。

追加した組織管理者ユーザのユーザ名、パスワードは忘れないように保管することを推奨します。

4.11. 拡張対話（Web検索）機能のセットアップ

本製品では、外部の検索エンジンサービスの検索結果を用いてLLMと対話するWeb検索機能を提供しています。Web検索機能はチャット画面とAPIで利用できます。詳細は「チャット画面利用ガイド」「拡張対話チュートリアル」をご確認ください。

- Web検索機能を使うには別途検索エンジンサービスとの契約が必要です。契約はお客様ご自身で行う必要があります。
- 利用可能な外部の検索エンジンサービスは、「Brave Search」「Tavily」です。
- 本機能はインターネット通信ができない閉域環境では使用できません。

Web検索機能をご利用の際は、別途それぞれの検索エンジンサービスの規約に従い、APIキーの払い出しを実施してください。

APIキー払い出し後、以下の手順を踏むことでWeb検索機能が利用可能になります。

1. /opt/nec/genai/config/genai.envを開きます。

```
vi /opt/nec/genai/config/genai.env
```

2. WEBSEARCH_ENGINE、WEBSEARCH_API_KEYを記載します。
   「Brave Search」の場合は”0”、「Tavily」の場合は”1”を記載してください。

```
WEBSEARCH_ENGINE=”0”
WEBSEARCH_API_KEY=<払い出した検索エンジンサービスのAPIキー>
```

3. genaiサービスを再起動します。

```
systemctl restart genai
```

4.12. APIキーの更新

Generative AI FW では、Keycloakによるユーザ認証だけでなく、API利用者に向けたAPIキーでの認証をサポートしています。初期のAPIキーから、必ずAPIキーを変更してください。

APIキーの更新は管理ポータル画面から行います。詳細は「管理ポータル操作ガイド（設定画面編）」の「APIセクションの説明」をご確認ください。

4.13. 新ユーザ、新規APIキーでの正常性確認

新規ユーザ、新規APIキーの情報が正常に動作するか管理ポータル画面、チャット画面、APIの観点で確認していきます。

正常性確認手順は「正常性確認ガイド」をご確認ください。

なお、チャット画面ではLLMをセットアップ完了しているため、対話ができることも合わせて確認してください。

拡張対話（Web検索）機能は正常性確認手順には含まれておりません。必要があればチャット画面利用ガイドを参照し問題なく動作するか確認してください。

4.14. 認証方式の検討

Generative AI FWでは複数の認証方式に対応しています。対応している認証方式と機能差異は以下の通りです。カスタム認証の詳細については「カスタム認証セットアップガイド」をご確認ください。

認証方式	ログインID	ユーザ作成時のパスワード設定	ログインパスワードの初期化方法	ユーザ・グループの同期機能
初期認証	メールアドレス	必須	管理者によるログインパスワードの強制初期化	無
カスタム認証（ユーザID認証）	ユーザー名	必須	管理者によるログインパスワードの強制初期化	無
カスタム認証（Entra ID）	EntraIDの設定に依存	不要	EntraID側で実施	無
カスタム認証（Active Directory）	ユーザー名	不要	Active Directory側で実施	有

デフォルトでは、初期認証に設定されています。「ユーザ・グループの同期機能」とはカスタム認証（Active Directory）のみ利用可能な機能でActive Directoryで作成したユーザ・グループを手動または自動によってGenerative AI FWのユーザ・グループ情報と同期させることができる機能です。カスタム認証への変更については「カスタム認証セットアップガイド」をご確認ください（ユーザ・グループの同期機能についても記載しています）。なお、ログインパスワード忘れた場合の対応方法を「管理者によるログインパスワードの強制初期化」で行う場合の操作方法は別紙の「運用ガイド」をご確認ください。

4.15. 初期ユーザの削除

最後に、出荷時の初期ユーザを削除します。以下の手順に従い、ユーザを削除してください。

本手順は必ず一人以上の組織管理者ユーザが存在し、管理ポータル画面にログインできることを確認してから実施してください。

1. 初期ユーザ以外の組織管理者ユーザで管理ポータルにログインします。
2. 左メニューからユーザを選択し、初期ユーザ（admin@example.com）の削除アイコンをクリックします。
3. 削除確認ダイアログが表示されますので、OKを押します。
4. 管理ポータルからログアウトします。
5. チャット画面にアクセスし、初期ユーザでログインできないことを確認します。

4.16. 管理者ユーザのパスワード変更

サーバの管理者ユーザadministratorの初期パスワードを変更します。管理者ユーザでログオンして以下のコマンドからパスワードを変更してください。

サーバの管理者ユーザのユーザ名、パスワードは忘れないように保管することを推奨します。

```
passwd
```