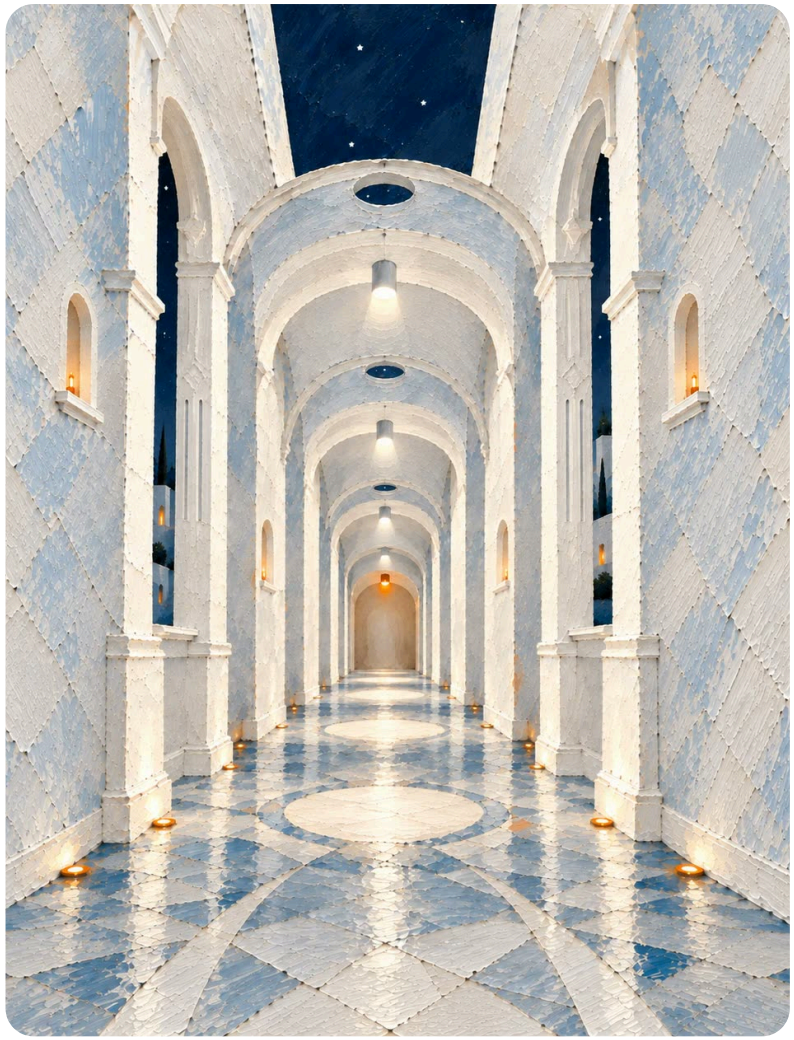
AIゲートウェイ

AIRouter

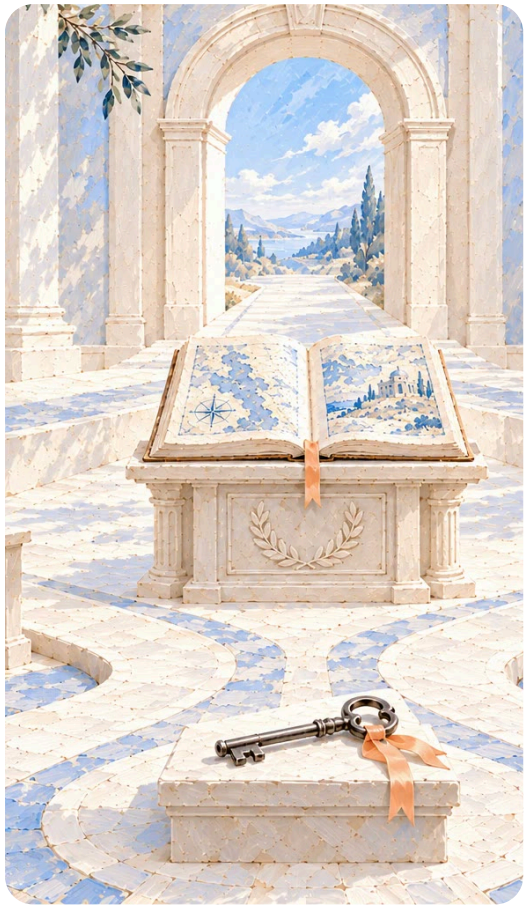
インテリジェントAIゲートウェイ・ソリューション

AIコストは請求書より先に把握し、
機密情報はプロバイダーに届く前にフィルタリングします。



このパンフレットは、
機能の一覧ではなく、AIモデルを
運用する企業が実際に口にする
7つの課題から始まります。

課題を1ページ、解決を1ページずつ、
順にお読みいただけます。



はじめに

はじめに　AIモデルを運用する企業の7つの課題　p.03

ソリューション　AIRouterとは何か、どこで課題を解決するのか　p.04

7つの課題と解決

課題 01　AIコストは請求書が届くまで分からない　p.05

課題 02　プロバイダーごとにキー・コンソール・契約が別々　p.07

課題 03　AIコストがどこで漏れているのか分からない　p.08

課題 04　個人情報がそのままAIに送られないか不安　p.10

課題 05　エージェントの行動を事故の後に説明できない　p.12

課題 06　モデルを1つ替えるにも再デプロイが必要　p.14

課題 07　プロバイダーが1社止まるとサービスも止まる　p.16

適用と導入

適用例　AIRouterの導入で変わるAI運用　p.17

導入ステップ　1つのサービスから始めて広げていく導入ステップ　p.19

お問い合わせ　デモ動画・導入のお問い合わせ　p.20

よく伺う課題

AIモデルを運用する企業がまず口にする7つの課題

複数のAIモデルをサービスに
組み込んだ瞬間から、
管理部門は統制を、
開発組織はスピードを求めます。

双方から繰り返し伺う声を
まとめました。



最も気になる課題をクリックすると、
AIRouterによる解決策の
ページへすぐに移動します。

課題 01	“AIのコストは、請求書が届くまで分かりません。”	予算統制 →
課題 02	“プロバイダーごとに、キーもコンソールも契約も別々です。”	プロバイダー統合 →
課題 03	“AIのコストがどこで漏れているのか分かりません。”	コスト削減 →
課題 04	“顧客の個人情報がそのままAIに出ていかないか心配です。”	個人情報 →
課題 05	“エージェントがしたことを、事故の後で説明できません。”	エージェント統制 →
課題 06	“モデルを1つ替えるにも、コードを直して再デプロイです。”	モデル切り替え →
課題 07	“プロバイダーが1社止まると、うちのサービスも止まります。”	障害対応 →

ひとつのゲートウェイ

AIRouterは、これらの課題を ひとつのゲートウェイで解決します。

インテリジェントAIゲートウェイ・ソリューション

プロバイダーのコンソールやサービスごとの個別開発に分散していた予算・セキュリティ・障害対応をひとつのゲートウェイに集約し、その記録をAIが分析して対処を提示します。



「課題」の下の番号をクリックすると、その課題へ移動します。

課題 01 · 予算統制



AIのコストがいくらなのか、請求書が届くまで分からないんです。

財務・購買の担当者が、真っ先に口にする話です。

現場ではこんなことが起きています

使った後で金額が分かります

AIのコストはリクエストの量に応じて発生し、金額は翌月の請求書で初めて分かります。

誰がどれだけ使ったのか分けられません

プロバイダーの請求書はアカウント単位のため、部門・プロジェクト別に別途集計が必要です。

止める手段がありません

あるサービスが急に大量に呼び出しても、予算を超える前に止める方法がありません。

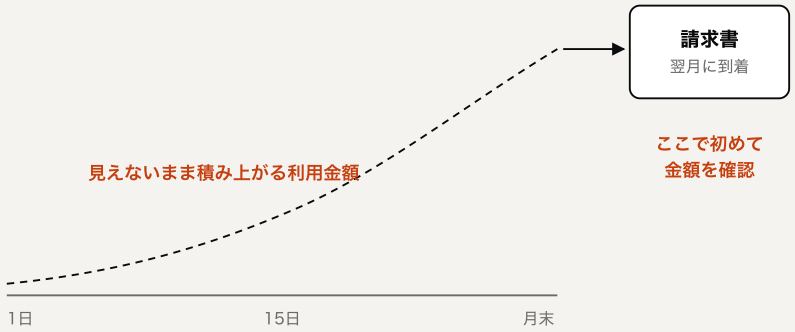
隠れた不安　“AI予算のコントロールを失うこと”

AIRouterの解決策を見る →

解決 01 · 事後の確認ではなく、事前の統制

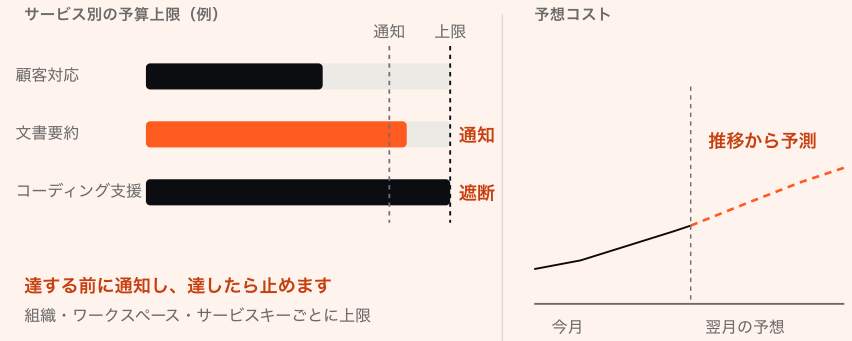
設定した上限に達する前に通知し、達したら呼び出しを止めます。

現在



利用から請求までの1か月間、コストが見えません。

AIRouter 導入後



利用した時点で単価で換算した見込み額を確認し、上限を超える呼び出しは送信前に拒否します。

階層ごとの予算上限

組織・ワークスペース・
サービスキーごとに、
使える金額を設定します。

達する前に通知、達したら遮断

上限に近づくと担当者に通知し、
上限に達すると呼び出しを止めます。

チーム・プロジェクト・モデル別に

複数プロバイダーの利用量と、
単価で換算した見込み額を1か所で、
必要な単位に分けて確認できます。

翌月のコストを予測

これまでの推移から、
翌週・翌月の予想コストを
事前に示します。

得られる価値

- ✓ 請求書の前に見込み額が分かります
- ✓ 設定した予算の範囲内で運用します
- ✓ 部門ごとの実績で予算を配分します

課題 02 · プロバイダー統合

“プロバイダーごとにキーもコンソールも契約も別々なんです。



サービスごとに個別に連携しています
アプリケーションごとにプロバイダーと
1つずつ接続し、管理する場所が増え続けます。

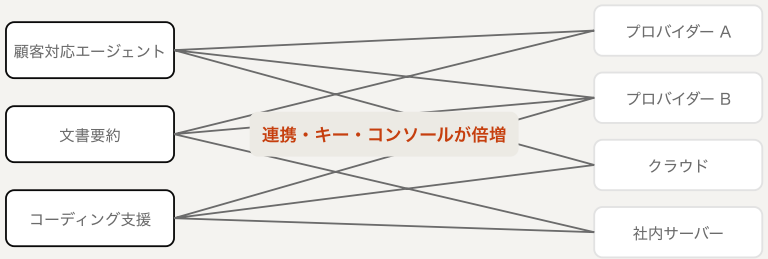
確認するコンソールもプロバイダーの数だけ
利用量・コスト・セキュリティ設定を、
プロバイダーのコンソールごとに別々に確認します。

隠れた不安
“プロバイダーが増えるほど、管理が追いつかなくなること”

解決 02 · ひとつのゲートウェイ

サービスキー1つで、複数プロバイダーのモデルを利用します。

現在



サービスとプロバイダーが増えるたびに、接続が倍々に増えます。

AIRouter



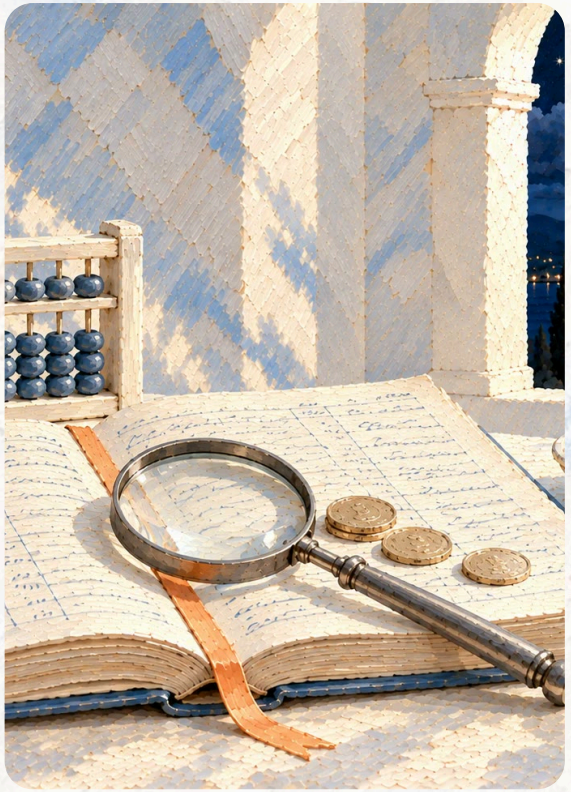
サービスは1か所にだけ接続し、プロバイダーはゲートウェイの背後で管理します。

得られる価値

- ✓ 既存のプロバイダー契約をそのまま利用できます
- ✓ プロバイダー別のコストを一目で把握できます

目次に戻る

課題 03 · コスト削減



AIのコストが
どこで漏れているか分からないんです。

AIサービスを運用する開発・財務の担当者が、ともに口にする話です。

現場ではこんなことが起きています

同じ内容を毎回計算し直しています

繰り返し送る長い指示文にキャッシュを使わず、
割引できるコストをそのまま払っています。

失敗したリクエストにも課金されます

再試行・失敗・途切れた応答のコストが、
請求書の中に埋もれて見えません。

簡単な作業にも最も高価なモデル

簡単な要約や分類にも、
最上位のモデルが使われています。

隠れた不安 **“削減できるコストがあるのに、見つけられないこと”**

AIRouterの解決策を見る →

解決 03 · 削減できるコストから

リクエスト記録を5分ごとにチェックし、削減できるコストとその方法をお知らせします。



9つのコストルール

キャッシュ未使用・再試行・失敗・
途切れた応答・長い会話を検出し、
削減見込み額を算出します。

自動プロンプトキャッシュ

繰り返される先頭部分に
キャッシュポイントを付け、
同じ会話は同じプロバイダーに送ります。

条件に合ったモデルへ

入力の長さ・ツールの利用・画像の
有無など、設定した条件に応じて
モデルを振り分けます。

数値はルールが、説明はAIが

金額と件数はリクエスト記録から
算出し、AIは原因と対処を
まとめます。

得られる価値

- ✓ 削減余地を見込み額で確認できます
- ✓ 改善箇所と改善方法を受け取れます
- ✓ 対処後の解消まで確認できます

課題 04 · 個人情報・機密情報



顧客の個人情報がそのままAIに出ていかないか心配です。

セキュリティ担当者が、AIサービスの決裁前に真っ先に尋ねる質問です。

現場ではこんなことが起きています

途中で確認する場所がありません

各サービスがプロバイダーを直接呼び出すため、何が外部に出ているのか確認できません。

読んだ文書内の指示に従ってしまいます

読み込んだ文書や検索結果に潜む指示が、エージェントの行動を変えることもあります。

応答にも機密情報が混ざります

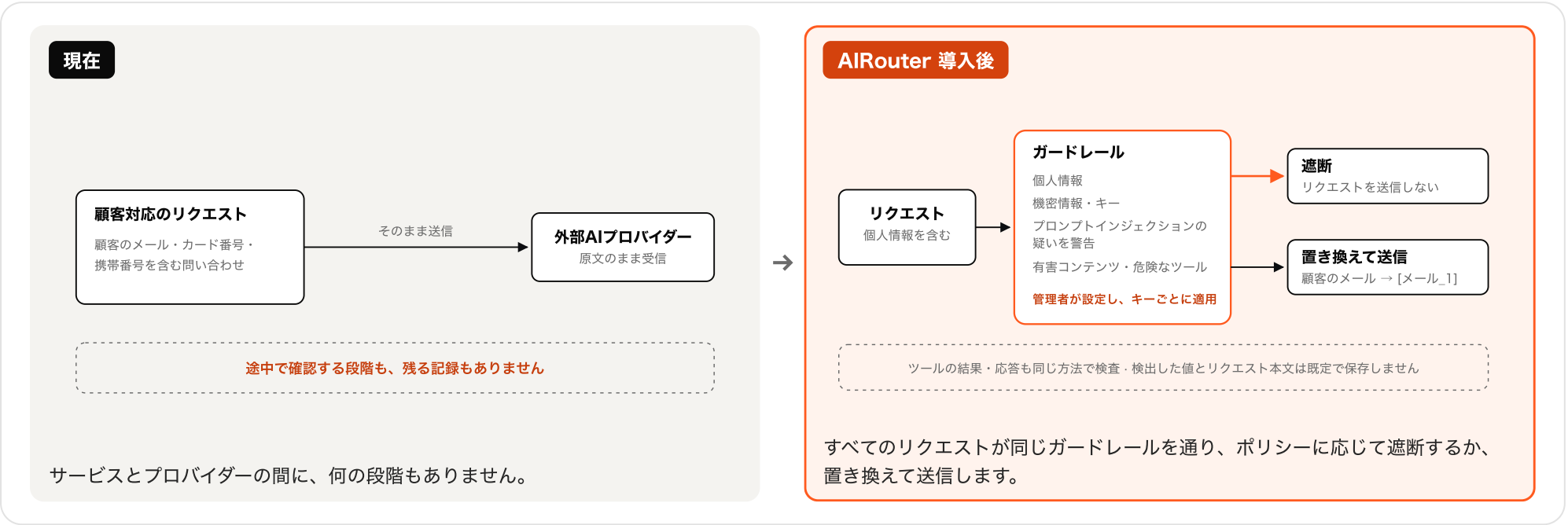
モデルの応答にキーやパスワードなどの機密情報が混ざっても、取り除けません。

隠れた不安 “流出事故の責任が、AI導入を決裁した人に返ってくること”

AIRouterの解決策を見る →

解決 04 · プロバイダーに届く前に

住民登録番号・カード番号・シークレットキーは、プロバイダーに届く前にフィルタリングします。



リクエストから応答まで検査

送信するリクエストだけでなく、エージェントが受け取ったツールの結果やモデルの応答まで検査します。

遮断するか、置き換えて送信

住民登録・外国人登録・パスポート・運転免許・口座の番号などは、既定で種別ラベルに置き換えて送信します。

開発者が解除できないポリシー

ガードレールは管理者が設定してサービスキーごとに適用し、呼び出す側では解除できません。

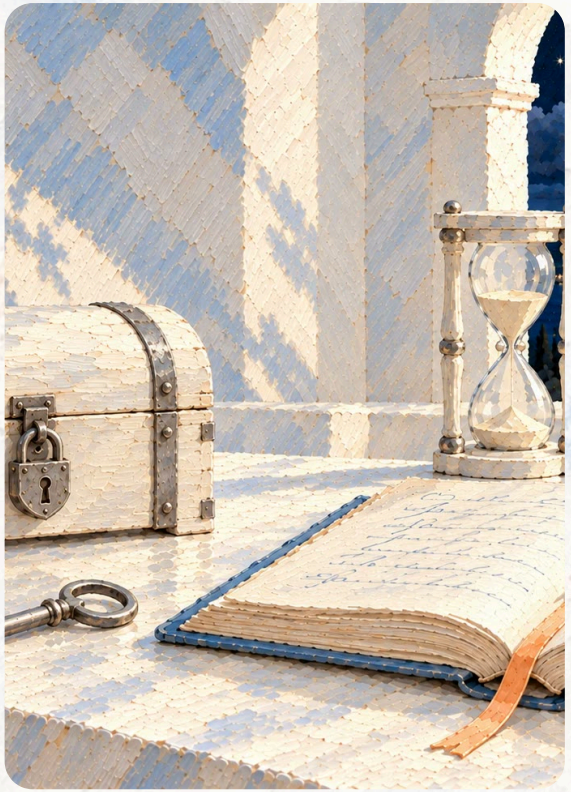
まず「記録のみ」、確認後に遮断

ルールを「記録のみ」で設定すると、遮断していればブロックされたリクエストの数を先に確認できます。

得られる価値

- ✓ センシティブ情報を送信前に止めます
- ✓ 業務を止めずにポリシーを試せます
- ✓ 全サービスに同じセキュリティ基準

課題 05 · エージェント統制・監査



エージェントが何をしたのか、事故が起きても説明できないんです。

エージェントを業務に導入しようとするIT・セキュリティ部門が尋ねる質問です。

現場ではこんなことが起きています

1つの作業が数十件の呼び出しに

作業1つがモデル・ツールの呼び出し数十件に広がり、1件ずつではリスクが見えません。

危険なツールもそのまま実行されます

メール送信・決済・削除・コマンド実行などのツール呼び出しを、途中で確認できません。

キーがどこで使われているか見えません

見慣れないIPからの利用や、休眠していたキーの再利用に気づけません。

隠れた不安 **“事故の後に、経緯をたどれる記録がないこと”**

AIRouterの解決策を見る →

解決 05 · エージェントの行動まで

危険なツール呼び出しは統制し、リスクの兆候は記録から先に見つけて知らせます。



危険なツール呼び出しの統制

送信・投稿・購入・削除・
コマンド実行などのツール呼び出しを、
ポリシーで統制します。

15のセキュリティルール

新しいIP・休眠キー・リクエスト急増・
拒否後の繰り返し試行などの兆候を
5分ごとに検出します。

指示に沿った行動か

危険なツールを呼んだセッションは、
ユーザーが最初に出した指示と
比べてAIが判定します。

即時停止と監査ログ

組織・ワークスペース・セッション
単位ですぐに止め、設定変更・
ログイン・本文の閲覧まで記録します。

得られる価値

- ✓ リスクの兆候に事故の前に気づけます
- ✓ 経緯を記録で説明できます
- ✓ 問題が起きたらすぐに止められます

課題 06 · モデル切り替え



モデルを1つ替えるにも、コードを直して再デプロイしています。

AIサービスを開発するチームが、最もよく口にする話です。

現場ではこんなことが起きています

モデル名がコードに埋め込まれています

新しいモデルが出るたびに、
修正・検証・デプロイを繰り返します。

元に戻すにもデプロイが必要です

切り替えたモデルに問題が起きると、
元に戻すにも再デプロイが必要です。

プロバイダーで呼び出し方が違います

プロバイダーを移ると、リクエスト形式や
応答の処理から合わせ直す必要があります。

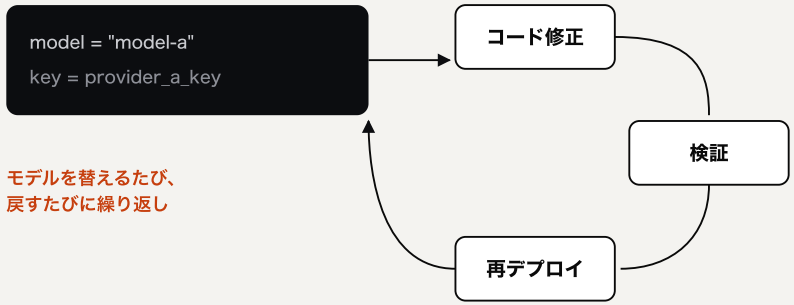
隠れた不安 “切り替えが怖くて、より良いモデルを使えないこと”

AIRouterの解決策を見る →

解決 06 · コードはそのまま、運用はコンソールで

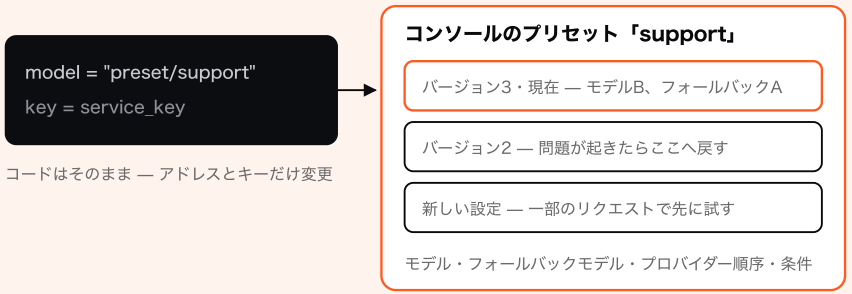
コードにはプリセット名だけを書き、モデルの切り替えと切り戻しはコンソールで行います。

現在



モデルが1つ変わるだけで、開発スケジュールが動きます。

AIRouter 導入後



モデルの運用はコンソールに移し、開発はサービスに集中します。（画面例）

アドレスとキーを変えるだけで接続

広く使われている2つのAI API
形式と互換性があり、既存のコードは
アドレスとキーの変更だけで
動作します。

ルーティングプリセット

モデル・フォールバックモデル・
プロバイダーの順序・条件を
コンソールに保存し、コードからは
名前だけを呼び出します。

バージョン管理と切り戻し

プリセットをバージョンで管理し、
問題が起きたらすぐに
以前のバージョンへ戻します。

一部のリクエストで先に試す

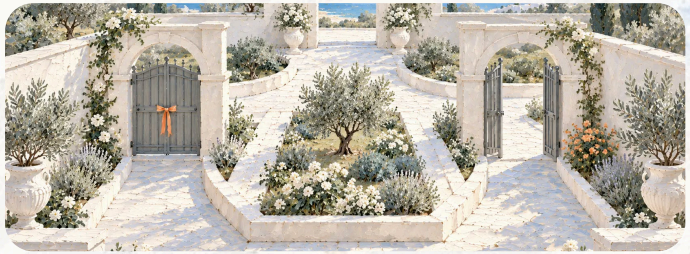
新しい設定を一部のリクエストにだけ
適用し、コストと応答速度を
確認してから広げます。

得られる価値

- ✓ モデルの切り替えにデプロイ不要
- ✓ 問題が起きたらすぐに戻せます
- ✓ より良いモデルを気軽に試せます

課題 07 · 障害対応

“プロバイダーが1社止まるだけで、うちのサービスも止まります。



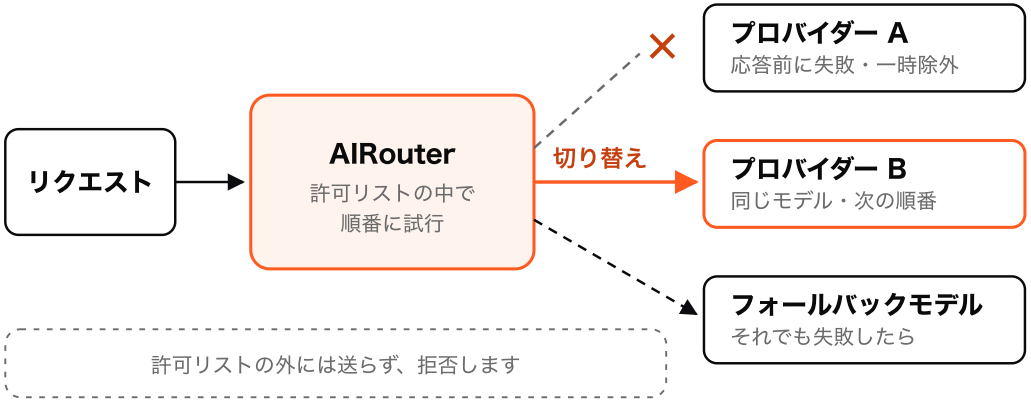
呼び出し上限の超過がそのままユーザーに
プロバイダーの呼び出し上限を超えると、
そのエラーがユーザーの画面まで届きます。

代替ロジックをサービスごとに作っています
再試行や別のプロバイダーへの切り替えを、
開発チームがサービスごとに自作しています。

隠れた不安
“AIプロバイダーの障害が、自社サービスの障害になること”

解決 07 · 許可された範囲内で切り替え

プロバイダーに問題が起きたら、許可済みの別プロバイダーへ切り替えます。



障害時の自動切り替え
応答前に失敗すると次のプロバイダーへ、さらにフォールバックモデルへ切り替えます。

失敗の多い先は一時除外
短時間に失敗が続いたプロバイダーには、しばらく送信しません。

ポリシー外には送りません
許可リスト・データ非保存・学習不使用の条件から外れる先には送らず、拒否します。

得られる価値
- ✓ 障害対応を各サービスで作る必要がありません
- ✓ 切り替えもポリシーの範囲内でのみ行います

目次に戻る

業務適用例

AIRouterを導入すると、AI運用はこう変わります。

エージェント開発・リリース前の点検

コストとセキュリティの問題に、リリース後に気づくことはありません

得られる価値 — **リリース前に**コストとセキュリティの問題を直せます。

現在は	AIRouterが担うこと	担当者が行うこと
顧客対応エージェントにかかるコストや呼び出すツールは、実際の顧客が使って初めて明らかになります。	開発環境のセッションを追跡し、キャッシュの無駄・危険なツール呼び出し・ガードレール判定を見つけ、原因と対処をまとめます。	指摘された箇所を直して再実行し、問題が解消したことを確認してからリリースします。

この場面を支える AIRouter: 解決 03 コスト分析 / 解決 05 ツール統制 / 解決 04 ガードレール

連携するもの: 開発ワークスペース / サービスキー / セッション・作業の記録

財務・部門別予算

部門別のAIコストを、手作業で分けません

得られる価値 — **部門別の実績をもとに**予算を配分します。

現在は	AIRouterが担うこと	担当者が行うこと
コンソールごとに利用量をダウンロードして部門別に分け、スプレッドシートで月次レポートを作っています。	ワークスペースごとの上限を守らせ、対話で依頼すれば表やグラフを含む月次レポートを作成します。	レポートを確認し、AIアシスタントが提案した予算調整を確かめて実行します。

この場面を支える AIRouter: 解決 01 予算統制 / 解決 02 1か所で集計 / 解決 03 コスト分析

連携するもの: ワークスペース / 予算上限 / AIアシスタント

業務適用の例です。貴社のサービスに合わせた範囲と効果は、導入のお問い合わせ（20ページ）で一緒に設計します。

業務適用例(続き)

同じ方法で、ほかの業務でも活躍します。

セキュリティ・規制業種

ポリシーをいきなり遮断にせず、まず試します

得られる価値 — **業務を止めずに**セキュリティポリシーを強化します。

現在は	AIRouterが担うこと	担当者が行うこと
遮断ルールを急に有効にすると業務が止まるのではと心配で、結局どのルールも有効にできません。	まず「記録のみ」でルールを設定し、遮断時にブロックされるリクエスト数をサービスキー別に示します。	影響を確認してから遮断に切り替え、監査ログをセキュリティ審査の資料として活用します。

この場面を支える AIRouter: 解決 04 ガードレール / 解決 05 セキュリティルール / 解決 01 予算統制

連携するもの: ガードレール / 監査ログ / 通知用Webhook

サービス運用・モデル切り替え

新しいモデルを、いきなり全体には適用しません

得られる価値 — **デプロイなしで**より良いモデルへ移行します。

現在は	AIRouterが担うこと	担当者が行うこと
より良いモデルが出ても全面的な切り替えが負担で、古いモデルを使い続けています。	プリセットの新バージョンを一部のリクエストで試し、障害時は許可された別のプロバイダーへ切り替えます。	リクエスト記録でコストと応答速度を比べてから、全体に広げるか元に戻します。

この場面を支える AIRouter: 解決 06 ルーティングプリセット / 解決 07 自動切り替え / 解決 03 コスト分析

連携するもの: ルーティングプリセット / リクエスト記録 / フォールバックモデル

業務適用の例です。貴社のサービスに合わせた範囲と効果は、導入のお問い合わせ（20ページ）で一緒に設計します。

導入ステップ

すべてのサービスを一度に移すことはしません。
まず1つのサービスを接続して確かめ、広げていきます。

1 2 3 4

現状を一緒に点検

どのサービスがどのプロバイダーを
使っているか、キーとコストが
どこで管理されているかを確認します。

得られる価値
散在するAI呼び出しとキーの現状把握

サービスを1つ接続

既存コードのアドレスとキーを変えて
サービスを1つ接続し、ガードレールは
まず「記録のみ」で設定します。

得られる価値
アドレスとキーを戻すだけで
元に戻せる、気軽な検証

ポリシーを有効化

予算上限とルーティングプリセットを
設定し、影響を確認したガードレール
ルールから遮断に切り替えます。

得られる価値
設定した予算・セキュリティ
基準の範囲内での運用

ほかのサービスへ展開

部門・プロジェクト別に
ワークスペースとサービスキーを分け、
同じポリシーのもとで広げていきます。

得られる価値
AI呼び出し全体を1か所で把握する運用

初回のご相談は、以下の一部があれば十分です

利用中のAIプロバイダー
契約中のAIプロバイダーと
クラウドの一覧

AIを呼び出すサービス
エージェント・アプリケーション・
社内ツールの一覧

月間のAIコスト規模
プロバイダー別の
おおよその利用金額

セキュリティ基準
外部送信を防ぐべき情報と
社内のセキュリティポリシー

SaaSで手軽に開始
サービスキーを受け取ってすぐに接続し、
既存のプロバイダー契約はそのまま利用します。

専用環境のご相談
セキュリティ要件の高い組織には、
お客様専用クラウド構成のご相談を承ります。

リクエスト本文は既定で保存しません
リクエスト本文は、組織が有効にした
場合のみ、決めた期間だけ保存します。

ご相談のご案内

まずは貴社の課題をお聞かせください。

AIを呼び出すサービスを1つお知らせいただければ、
そのサービスを基準に、コストが漏れている箇所・
外部に出ていくセンシティブ情報・障害時の切り替え経路を
デモでご説明します。資料は19ページの一覧の一部で十分です。
導入費用の算定基準もあわせてご案内します。
先に概要をご覧になりたい場合は、デモ動画をご確認ください。

導入のお問い合わせ →

▶ デモ動画を見る



お問い合わせの際は、最も急ぎの課題を1つお知らせください

予算統制 / プロバイダー統合 / コスト削減 / 個人情報 / エージェント統制 / モデル切り替え / 障害対応

メール
sales@vicsnc.com

ウェブサイト
vicsnc.com