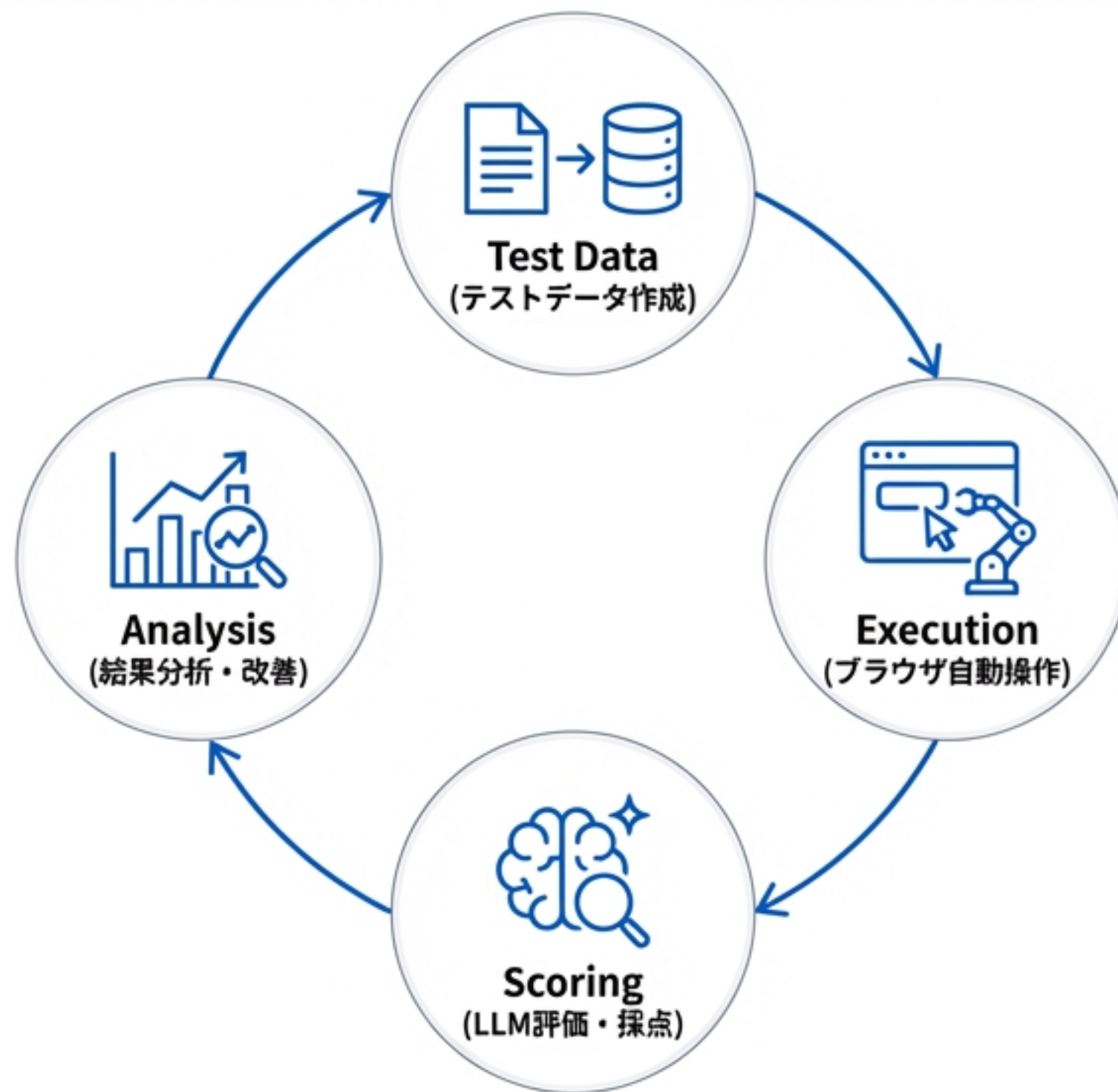


RAG汎用テストツール：操作マニュアルとワークフロー

PlaywrightとLLMスコアリングによるチャットボット精度評価の自動化



RAGチャットボットの回答精度を客観的かつ自動的に評価する

Core Value

手動テストの課題：時間がかかる、評価が主観的、スケールしない。

本ツールの解決策：Playwrightを用いた完全自動実行と、多次元的なスコアリングシステム。

 Dataset Management (Excel/CSV/JSON)

 RAG Connection Settings

 Automated Test Execution

 Score Evaluation & Analysis

Supported Browsers



対象ユーザー：QAエンジニア、LLM開発者、プロジェクトマネージャー

5つの指標を用いた多角的な精度測定

Lexical Matching (単語の一致)		
ROUGE-1	単語単位の一一致度	Range: 0.0 – 1.0
ROUGE-2	2単語連続の一一致度 (Bi-gram)	Range: 0.0 – 1.0
ROUGE-L	最長共通部分列の一一致度	Range: 0.0 – 1.0
BLEU	n-gram精度 (機械翻訳評価指標)	Range: 0.0 – 1.0
Meaning Matching (意味の類似)		
Semantic	LLMを使用した意味的類似度判定	Range: 0.0 – 1.0

RAGの文脈理解に最適な**最重要指標**
(Most critical metric for RAG context)

アカウントへのアクセスとダッシュボードの構成

Noto Sans JP Regular で JetBrains Mono に、エアカウントやデータセット数の場合、上からダッシュボード構成を記載されること、詳細は接面にファンコートの JetBrains JetBrains Mono です。

Access URL:

[https://\[Server-Address\]/rag-tester/](https://[Server-Address]/rag-tester/)

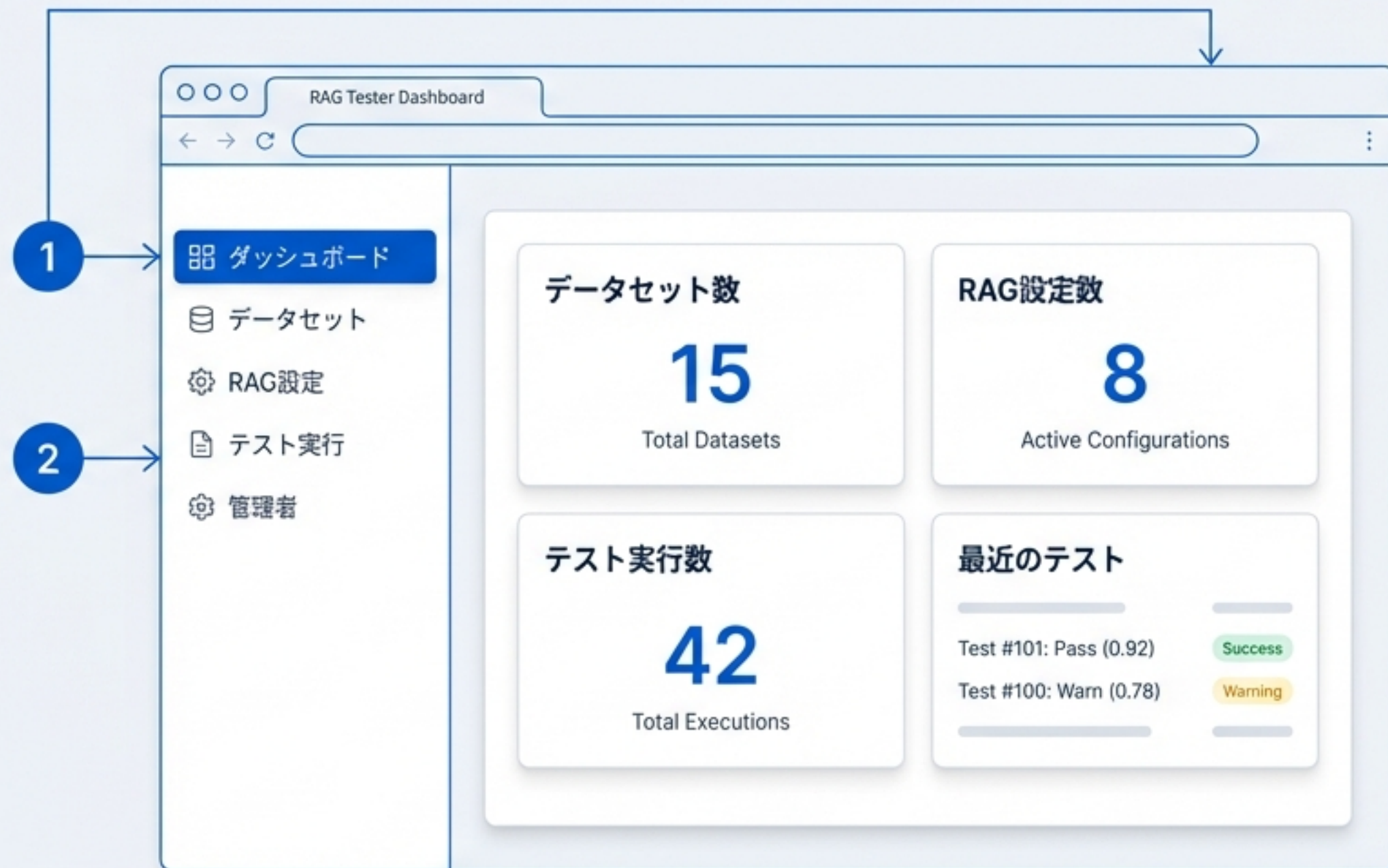
Login:

ユーザー名・パスワードを入力

Registration Rules:

User: 3-50 chars (Alphanumeric)

Pass: 8+ chars (Mix of upper/lower/num/symbol)



テストの基準となるデータセット（正解データ）の定義

Supported Formats



Excel (.xlsx)



CSV (.csv)

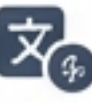


JSON (.json)

Requirements

 Max Size: 10MB

JetBrains Mono

 Encoding: UTF-8 (for CSV)

JetBrains Mono

	A	B	C	D
1	question ★ Must Have	answer ★ Must Have	category	tags
2	御社の営業時間は？	平日の9:00から18:00です。	Basic Info	hours,contact
2	パスワードを忘れました	ログイン画面の『パスワード再設定』から手続きしてください。	Support	login,trouble

AIによるクエリバリエーション生成でテストを強化する

- **Problem:**

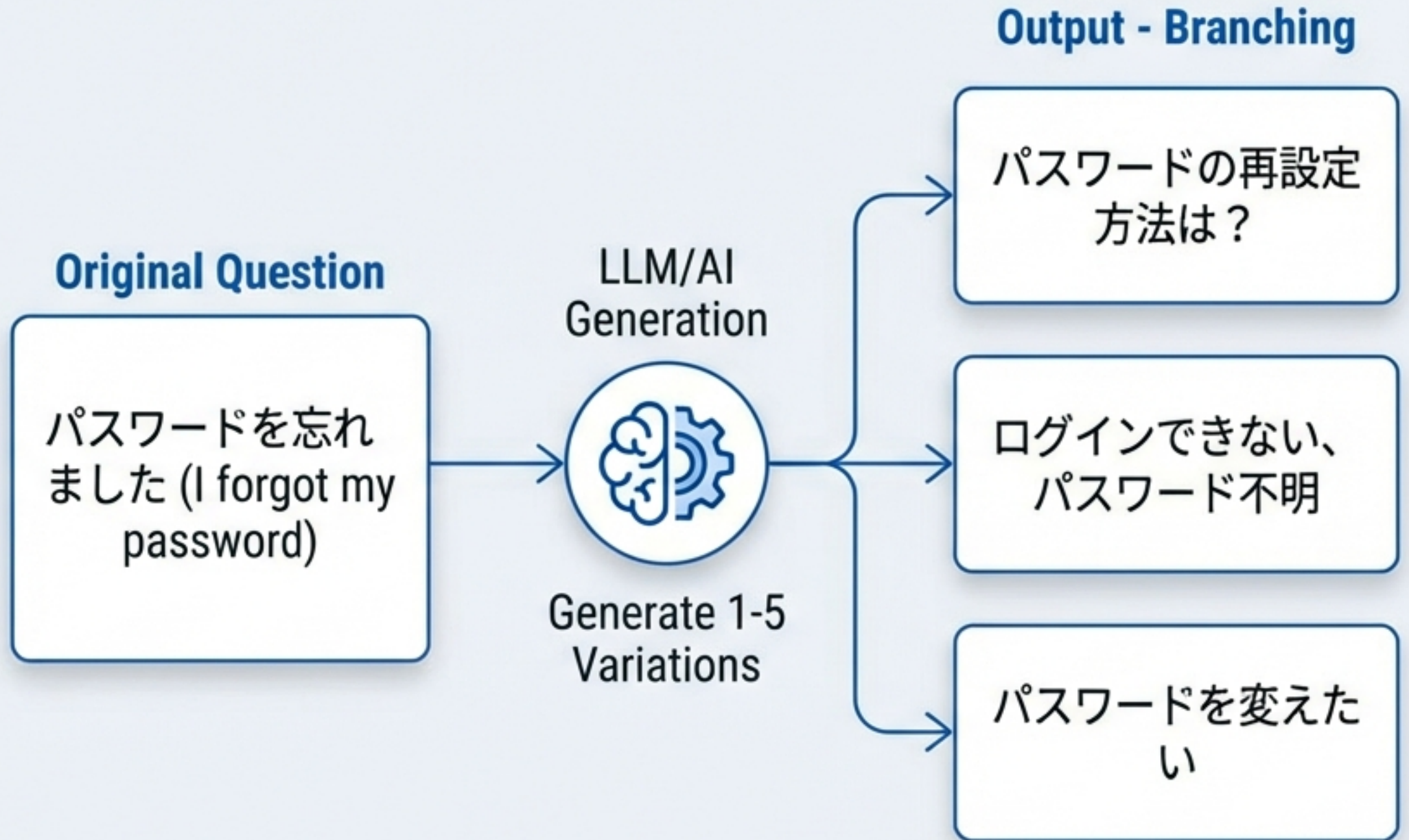
ユーザーは同じ質問を異なる言い回しで行う (Users phrase questions differently).

- **Action:**

データセット詳細画面 →
『クエリバリエーション生成』
をクリック

- **Benefit:**

手動での作成不要でテストカバレッジを拡大
(Expand test coverage without manual effort).



ターゲットRAGアプリケーションへの接続設定

Menu Path:

RAG設定 > 新規作成

Tip:

Botの応答が遅い場合は
Wait Timeを30秒以上
に設定してください
(Increase wait time if
bot is slow).

RAG設定 (New Creation)

Name (名前):

HR Support Bot v1

URL:

https://internal-rag.app

Wait Time (回答待機時間):

15

seconds (Default)

Authentication (認証):

- ☐ None
- ☒ Basic Auth
- ☐ Bearer Token
- ☐ Cookie

User (ユーザー):

Password (パスワード):

.....

ブラウザ操作のためのCSSセレクター特定



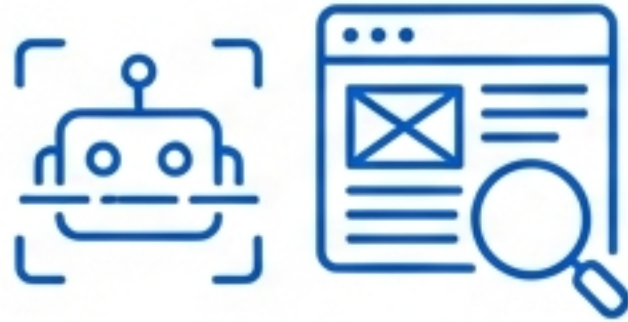
Streamlit Specific Tip

Streamlitアプリでは以下の属性を使用してください:

```
[aria-label="Chat message from assistant"][data-testid="stMarkdownContainer"]
```

セレクターの自動検出と接続検証

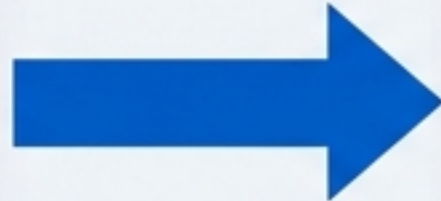
Auto-Detect (自動検出)



LLM analyzes page structure.



Auto-detection is not perfect.
Always verify.



Verification (検証)



Click “Verify” button.

1. Access App
2. Send “Test” query
3. Capture response



Verified (検証済み)

テストの実行とリアルタイムモニタリング

Test Configuration (テスト設定)

Select Dataset + RAG Setting.

****Query Variations****

(クエリバリエーション使用)

ON



Live Execution Panel (実行パネル)

 Running... (実行中)

60 / 100 Queries

Success Green

55

✓ Success

Warning Yellow

4

✗ Fail

Alert Red

1

! Error

■ Stop

↺ Re-run

合否判定ロジックとスコア基準

PASS (合格)

**ROUGE-1 ≥ 0.7
AND
Semantic ≥ 0.8**

High lexical and
meaning accuracy.

WARN (警告)

**ROUGE-1 ≥ 0.4
OR
Semantic ≥ 0.6**

Moderate accuracy.

FAIL (不合格)

Below WARN
thresholds

Low accuracy.

Error (Grey): Failed to retrieve answer.

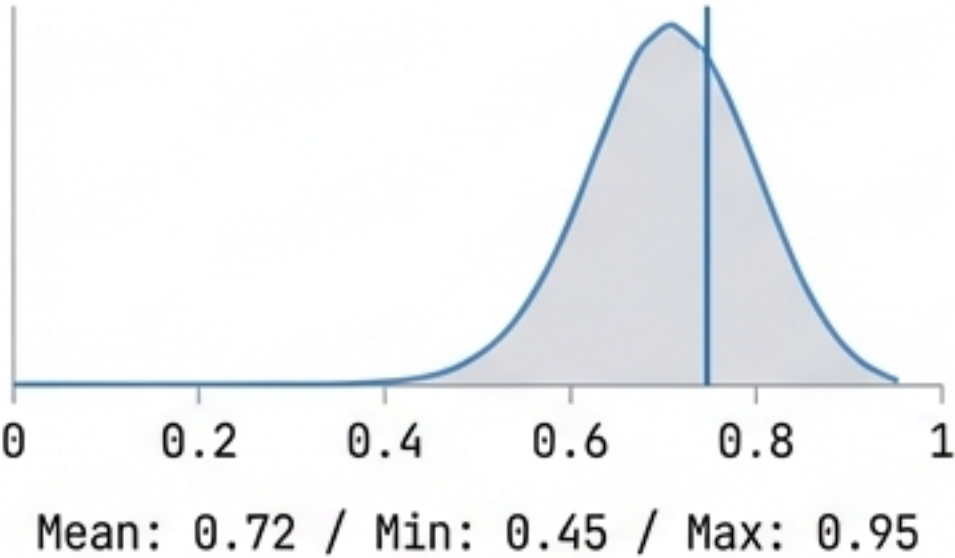
詳細結果の分析とトレンドの把握

Detailed View (詳細結果)

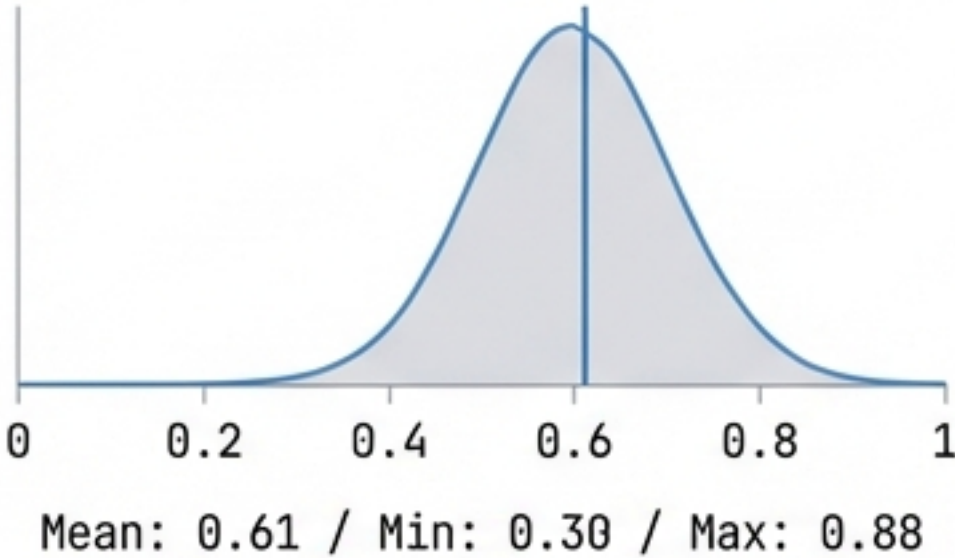
Query (クエリ)	Expected (期待値)	Actual (実測値)	Scores (スコア)
営業時間？	9:00-18:00	平日9時から6時	R-1: 0.75 Sem: 0.92 Pass (合格)

Statistical Analysis (統計分析)

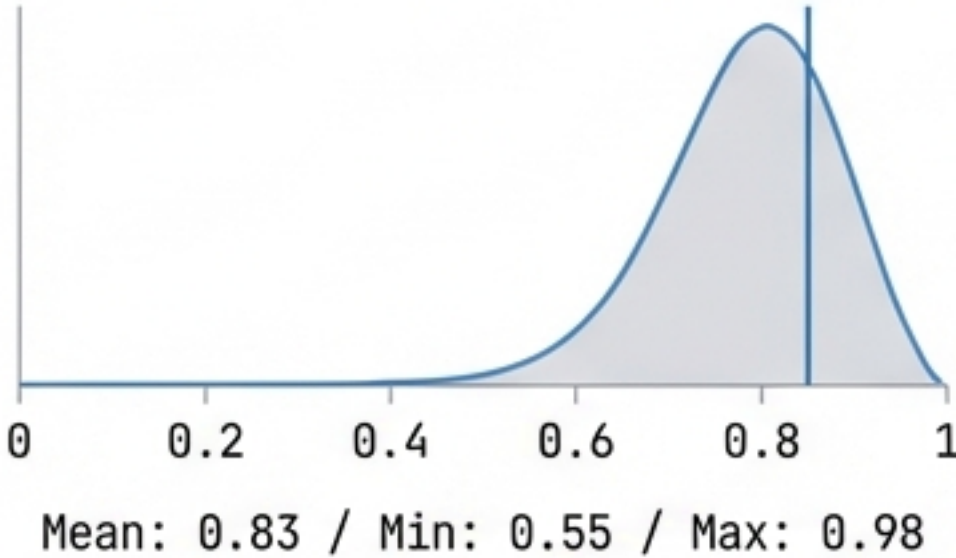
ROUGE Distribution (ROUGE分布)



BLEU Distribution (BLEU分布)



Semantic Distribution (意味的分布)



トラブルシューティング：接続とセレクター



**Timeout
Error**

Fix: Increase Wait
Time (>30s).



**Element
Not Found**

Fix: Use F12 Dev
Tools.
Re-run Verification.



**Same
Answer Loop**

Fix: Target *last*
message bubble using
aria-label.

トラブルシューティング：データとアカウント

データアップロードの問題



- ✓ Format: Excel, CSV, JSON
- ✓ Columns: question, answer (Required)
- ✓ Encoding: UTF-8
- ✓ Size: < 10MB

ログインの問題



- ✓ Check Username/Password
- ✓ Check Caps Lock
- ✓ Contact Admin for reset

リファレンスと重要事項

Keyboard Shortcuts

Key Combination	Action
Ctrl + Enter JetBrains Mono	Submit Form Slate Grey #344563
Esc JetBrains Mono	Close Dialog Noto Sans JP Regular, Slate Grey #344563

Glossary

RAG: Retrieval-Augmented Generation

Playwright: Browser Automation

Semantic: Meaning Similarity

“Regular automated testing is the key to maintaining RAG reliability.”

Contact Administrator for system support.

Version 1.0 (2026/02/18)
Noto Sans JP Regular, JetBrains Mono