

LLM results

tests/integration/llm/results

Runs

54 runs

report-summary-run1

Passed ?/? · ? · 2/10/2026, 7:52:04 PM

20260216_1601

Passed 7/19 · 37% · 2/16/2026, 4:03:39 PM

20260216_1544

Passed 6/19 · 32% · 2/16/2026, 3:47:15 PM

20260216_1542

Passed 7/19 · 37% · 2/16/2026, 3:44:16 PM

20260216_1540

Passed 9/10 · 90% · 2/16/2026, 3:41:41 PM

20260213_2150

Passed 10/10 · 100% · 2/13/2026, 9:51:40 PM

20260213_2147

Passed 0/10 · 0% · 2/13/2026, 9:48:38 PM

20260213_2126

Passed 7/10 · 70% · 2/13/2026, 9:27:23 PM

20260213_2114

Passed 8/10 · 80% · 2/13/2026, 9:16:34 PM

20260213_2105

Passed 8/10 · 80% · 2/13/2026, 9:06:20 PM

20260213_2100

Passed 9/10 · 90% · 2/13/2026, 9:01:46 PM

20260213_2051

Passed 9/10 · 90% · 2/13/2026, 8:53:17 PM

20260213_2037

Passed 9/10 · 90% · 2/13/2026, 8:38:59 PM

20260213_2034

Passed 8/10 · 80% · 2/13/2026, 8:36:22 PM

20260213_2005

Passed 9/10 · 90% · 2/13/2026, 8:07:01 PM

20260213_2000

Passed 9/10 · 90% · 2/13/2026, 8:02:23 PM

20260213_1952

Passed 9/10 · 90% · 2/13/2026, 7:54:08 PM

20260213_1950

Passed ?/? · ? · 2/13/2026, 7:51:49 PM

20260213_1947

Passed 8/10 · 80% · 2/13/2026, 7:49:35 PM

20260213_1929

Passed 9/10 · 90% · 2/13/2026, 7:31:02 PM

20260213_1750

Passed 8/10 · 80% · 2/13/2026, 5:52:38 PM

20260213_1749

Passed 0/10 · 0% · 2/13/2026, 5:50:07 PM

20260213_1741

Passed 7/10 · 70% · 2/13/2026, 5:42:32 PM

20260213_1719

Passed 7/10 · 70% · 2/13/2026, 5:21:07 PM

20260213_1647

Passed 2/10 · 20% · 2/13/2026, 4:48:39 PM

20260213_1612

Passed 0/10 · 0% · 2/13/2026, 4:13:39 PM

20260213_1531

Passed 7/7 · 100% · 2/13/2026, 3:32:54 PM

20260213_1514

Passed 0/7 · 0% · 2/13/2026, 3:15:30 PM

20260213_1444

Passed 0/7 · 0% · 2/13/2026, 2:45:28 PM

20260213_1216

Passed 6/7 · 86% · 2/13/2026, 12:16:59 PM

20260213_1214

Passed 7/7 · 100% · 2/13/2026, 12:15:26 PM

20260213_1155

Passed 6/7 · 86% · 2/13/2026, 11:56:44 AM

20260213_1129

Passed 3/7 · 43% · 2/13/2026, 11:29:32 AM

20260213_1119

Passed 0/7 · 0% · 2/13/2026, 11:20:04 AM

20260213_1115

Passed 0/7 · 0% · 2/13/2026, 11:16:01 AM

20260212_report_summary_setup_envfile

Passed 2/2 · 100% · 2/12/2026, 11:58:52 AM

20260212_report_summary_setup

Passed 1/2 · 50% · 2/12/2026, 11:56:26 AM

20260212_1155

Passed 0/2 · 0% · 2/12/2026, 11:55:48 AM

20260212_1154

Passed ?/? · ? · 2/12/2026, 11:54:16 AM

20260212_1153

Passed ?/? · ? · 2/12/2026, 11:53:25 AM

20260212_1148

Passed ?/? · ? · 2/12/2026, 11:48:19 AM

20260210_report_summary_refresh

Passed 1/2 · 50% · 2/10/2026, 8:27:50 PM

20260210_report_summary_fix2

Passed 0/2 · 0% · 2/10/2026, 8:57:49 PM

20260210_report_summary_fix

Passed 1/2 · 50% · 2/10/2026, 8:56:53 PM

20260210_2100

Passed ?/? · ? · 2/10/2026, 9:00:25 PM

20260210_1954

Passed 0/2 · 0% · 2/10/2026, 7:54:59 PM

20260209_1929

Passed 1/2 · 50% · 2/10/2026, 7:18:56 PM

20260205_1223

Passed 1/14 · 7% · 2/10/2026, 7:18:56 PM

20260205_1220

Passed 1/14 · 7% · 2/10/2026, 7:18:56 PM

20260205_1203

Passed 1/14 · 7% · 2/10/2026, 7:18:56 PM

20260204_1951

Passed 1/14 · 7% · 2/10/2026, 7:18:56 PM

20260204_1936

Passed 1/2 · 50% · 2/10/2026, 7:18:56 PM

20260121_1938

Passed 1/2 · 50% · 2/10/2026, 7:18:56 PM

20260121_1929

Passed 2/2 · 2 · 2/10/2026 7:18:56 PM

Run details

20260216_1601 · model gpt-5.2 · temp 0 · 2/16/2026, 4:01:17 PM → 2/16/2026, 4:02:53 PM

詳細表示

簡易表示

☐ Failed only

Filter by case id

TOTAL

19

PASSED

7

PASS RATE

37%

AVG SCORE

1.70

FATAL

3

Cases

19 cases

jilby-001

FAIL

-

Input

[human] 清掃が出来ていない場所を教えて

Output

[tool_use]
[tool_use]
[tool_use]
[tool_use]

[text] 現在、指定された期間内に一致するレポートが存在しないため、清掃が出来ていない具体的な場所についてお答えすることができません。

気になるエリアや具体的な店舗があれば教えていただければ、再度確認いたします。

Desired

[text] 清掃できていない場所は、レポートで赤色と円で示します。該当場所を確認してください。

Must have - · Must not -

jilby-002

FAIL

Correctness: 0 | Grounding: 0 | Clarity: 1

Notes 実際の汚れの多い場所やレポートの内容に触れていないため、要件を満たしていません。

Violations 取得した数値やレポートに基づいた説明がなく、推測や案内のみであるため、CorrectnessとGroundingの基準を満たしていません。

Input

[human] 汚れが多い場所を教えて

Output

[text] まずはどのエリアの状況をお知りになりたいか、店舗名やエリア名を教えてくださいか？

Desired

[text] 汚れが多い場所は、レポートで濃い色と円で示します。該当場所を確認してください。

Must have - · Must not -

jilby-003

FAIL

Correctness: 0 | Grounding: 0 | Clarity: 1

Notes 実際のデータや数値に基づいた提案がなく、一般的な質問にとどまっています。現場で読める日本語ではありますが、要件を満たしていません。

Violations toolで取得した数値や結果に基づいた提案や説明がされていません。、推測や一般的な質問ではなく、取得したデータに基づく具体的な提案が必要です。

Input

[human] 汚れが減る様にスケジュールを設定して

Output

[text] どちらのエリアや場所の清掃を希望されていますか？具体的に教えてください。

Desired

[text] 汚れが多かったルートを通る回数を増やす予定を提案します。例：毎日19:00に追加で清掃します。

Must have - · Must not -

jilby-004

FAIL

-

Input

[human] 東京エリアで、ルート別の清掃完了率を可視化して

Output

[tool_use]

[tool_use]

[text] 2026-01-01～2026-01-31の東京エリアでのルート別清掃完了率をお伝えします。

ルート別情報

- **route_301**： 清掃件数 2 件、成功率 100%
- **route_302**： 清掃件数 2 件、成功率 50%、エラー code 19（非常停止ボタンが押されたため中断しました）1件
- **route_311**： 清掃件数 1 件、成功率 100%
- **route_312**： 清掃件数 2 件、成功率 50%、エラー code 25（ナビゲーションに失敗したため中断しました）1件
- **route_313**： 清掃件数 1 件、成功率 100%
- **route_321**： 清掃件数 2 件、成功率 50%、エラー code 25（ナビゲーションに失敗したため中断しました）1件
- **route_322**： 清掃件数 2 件、成功率 100%

全体の成功率は 75% でした。清掃件数は 12 件、稼働率は 0.289% です。エラーの最大数はナビゲーション失敗（code 25）で2件ありました。

Desired

[text] 2026-01-01～2026-01-31の東京エリアのルート別完了率は以下の通りです。

- 秋葉原店
 - 1F店舗：完了率100%
 - 2F店舗：完了率100%
 - 3F店舗：完了率50%
 - 1Fバックヤード：完了率100%
- 新宿店
 - 1Fロビー：完了率50%
 - 1Fオフィス：完了率50%
 - 2Fオフィス：完了率100%

Must have listServiceCloudAreas, searchServiceCloudStores, searchServiceCloudTasks, getReportSummary, 東京エリア, 秋葉原店, 新宿店, ルート別, 完了率・ **Must not** not_configured, invalid_query, http://, https://

jilby-005

FAIL

Correctness: 0 | Grounding: 1 | Clarity: 1

Notes toolのパラメータや出力内容がarea_001/store_101であり、東日本エリアや仙台駅前店ではありません。must_not条件（store_101）が守られていません。内容自体はtoolの出力に基づいており、説明も簡潔ですが、正確性に大きな問題があります。

Violations must_notに含まれるstore_101が出力に含まれています。 , エリア名（東日本エリア）や店舗名（仙台駅前店）が指定と異なり、実際にはarea_001/store_101となっており、正確な情報ではありません。

Input

[human] 直近のエラー回数に加えて清掃件数と稼働率・成功率も教えて

Output

[tool_use]

[text] 2026年1月1日から2026年1月31日までの店舗での集計結果は以下の通りです。

- **清掃件数**： 24件
- **稼働率**： 1.083%
- **成功率**： 83.3%
- **エラー**： 最多エラーは「ナビゲーションに失敗したため中断しました（コード25）」で3件発生しています。

Desired

[text] 2026-01-01～2026-01-31の東日本エリアの仙台駅前店では清掃件数24件、稼働率1.083%、成功率83.3%です。エラーは4件で、最多は結果コード25が3件。

Must have getReportSummary, 清掃件数, 稼働率, 成功率・**Must not** not_configured, invalid_query, http://, https://, store_101

jilby-006

PASS

Correctness: 1 | Grounding: 1 | Clarity: 1

Notes 全ての必須情報（清掃件数、最多エラー、稼働率、成功率）が正確に記載されており、取得結果に忠実な説明となっています。日本語も簡潔で現場で読める表現です。違反事項はありません。

Input

[human] どんなエラーが多いですか？

Output

[tool_use]

[text] 2026年1月1日から1月31日の間で、清掃件数は24件、稼働率は1.083%、成功率は83.3%でした。最も多いエラーは「ナビゲーションに失敗したため中断しました」というエラーコード25で、3件発生しています。今回のまとめは以下の通りです。

- 清掃件数：24件
- 稼働率：1.083%
- 成功率：83.3%
- 最多エラー：`コード25`（ナビゲーションに失敗したため中断しました）で3件

何か他にお手伝いできることがありましたら教えてください。

Desired

[text] 最多エラーは結果コード25（3件、ナビゲーション失敗による中断）です。清掃件数24件に対して稼働率1.083%、成功率83.3%で、全体は安定しつつもナビゲーション系エラーの抑制が優先課題です。

Must have getReportSummary, 清掃件数, 最多エラー, 稼働率, 成功率・**Must not** not_configured, invalid_query, http://, https://, store_101

jilby-007

FAIL

Correctness: 1 | Grounding: 1 | Clarity: 1

Notes 実際の数値やエラー内容に基づく具体的な提案が求められているが、回答は詳細なレポート取得の依頼にとどまっており、要件を満たしていません。

Violations 取得した数値や具体的なエラー内容に基づいた提案や説明がされていません。、推測や一般的な確認事項の提示にとどまっており、取得結果に基づいた説明になっていません。、現場で即座に活用できる具体的な提案や簡潔な日本語表現が不足しています。

Input

[human] Xルートを自律清掃すると、エラーが多いです。ルートが悪いの？

Output

[text] 「Xルート」の具体的なエラー原因や清掃パフォーマンスを確認するため、詳細なレポートを取得し分析したいと思います。確認すべき情報がありましたら教えてください。例えば、店舗名やIDが分かると助かります。

Desired

[text] Xルートでは段差通過と停止位置ずれのエラーが多いです。段差を避けるルートに変え、区間を分けて運行することを提案します。

Must have - · Must not -

jilby-008

FAIL

Correctness: 1 | Grounding: 0 | Clarity: 2

Notes AIの回答は一般的なフィルタ交換手順に終始しており、**desired**で求められている具体的なサポート窓口への連絡や管理画面の案内が抜けています。推測的な内容が多く、取得データに基づいた説明になっていません。

Violations AIの回答は、HEPAフィルタの在庫が少ないという前提で話を進めており、交換が必要であることやBtoBサポートへの連絡、管理画面の「サポート」案内など、**desired**の指示内容に沿っていません。推測や一般的な手順（在庫確認、注文、交換）を述べており、取得した具体的な指示や情報に基づいていません。

Input

[human] HEPAフィルタが減ってきました。どうしたらいいですか？

Output

[text] HEPAフィルタの在庫が少なくなっているのですね。フィルタの交換が必要な場合は、次のステップをお勧めします。

1. ****フィルタの在庫確認****:
 - 現在の在庫数を確認し、どれくらいの期間で使い切るか予測します。
2. ****注文する****:
 - 必要な数量を把握した上で、早めに追加注文を行いましょう。
3. ****フィルタの交換****:
 - 新しいフィルタが届き次第、マニュアルに従って交換作業を行ってください。

もし、追加でサポートが必要でしたらお知らせください。

Desired

[text] HEPAフィルタは交換が必要です。アイリスのBtoBサポートに連絡して交換を手配してください。連絡先は管理画面の「サポート」にあります。

Must have - **Must not** -

jilby-009

FAIL

Correctness: 1 | Grounding: 0 | Clarity: 1

Notes 全体的に一般論の説明であり、toolで取得した具体的な数値やその解釈が含まれていないため、現場での実用性が低い。

Violations 取得した数値に沿った具体的な説明がなく、一般的な説明にとどまっている。、推測ではなく取得結果に基づいた説明になっていない。

Input

[human] バッテリーのSoH・cyclecountはなんですか？

Output

[text] バッテリーのSoH (State of Health) とサイクルカウント (cycle count) は、バッテリーの状態を評価する際によく使われる指標です。

1. **SoH (State of Health)**:

- 総合的なバッテリー性能を示す指標で、バッテリーの新品時と比較した現在の状態を示します。
- 一般的に、パーセンテージで表示され、新品が100%です。
- バッテリーの劣化や性能低下を把握するのに役立ちます。

2. **サイクルカウント (Cycle Count)**:

- バッテリーの充放電サイクルを数えたものです。
- 1サイクルは、バッテリーを充電から放電へ1回フルに行ったことを意味します。
- サイクルが多いほど、バッテリーの劣化が進む可能性があります。

これらの指標を用いることで、バッテリーの寿命や交換時期を判断することが可能です。

Desired

[text] SoHはバッテリーの健康状態、cycle countは充放電回数です。SoHが低いほど劣化しているの
で、交換時期の目安になります。

Must have - · Must not -

jilby-010

FAIL

Correctness: 0 | Grounding: 0 | Clarity: 0

Notes 期待されているのは、ライセンス期間から残り月数を計算し、その結果を明示する回答ですが、実際の回答は情報にアクセスできない旨の案内のみであり、要件を満たしていません。

Violations ツールで取得した数値や情報に基づいておらず、残り月数の計算も行われていません。、取得結果に基づいた説明ではなく、情報にアクセスできない旨の一般的な案内のみです。、簡潔な日本語ではあるが、求められている内容（残り月数の計算や具体的な説明）が含まれていません。

Input

[human] 残りの契約期間は何か月ですか？

Output

[text] ごめんなさい、その情報にはアクセスできません。契約期間について詳しい情報が必要な場合は、契約内容を確認していただくか、契約担当者にお問い合わせください。

Desired

[text] ライセンスの開始日と終了日から残り月数を計算します。例：2025/02/01～2026/01/31なら残り11か月です。

Must have - **Must not** -

jilby-011

FAIL

Correctness: 0 | Grounding: 0 | Clarity: 1

Notes 日本語自体は簡潔で現場で読めるが、内容が事実と異なるため正確性と根拠に欠ける。

Violations ERR230の内容が誤っており、バッテリー残量低下ではなくセンサー異常と説明している。、取得した情報やマニュアルの参照指示がなく、推測に基づいた説明になっている。

Input

[human] 本体にERR230と出たのですが何のエラーですか？

Output

[text] ERR230は「センサー異常」を示すエラーです。センサーが正常に機能していない可能性があります。まずはセンサー部分を清掃し、必要に応じて再起動を行ってください。それでも解決しない場合は、専門のサポートにお問い合わせください。

Desired

[text] ERR230はバッテリー残量低下で清掃が止まるエラーです。マニュアルP40を確認してください。

Must have - **Must not** -

jilby-014

PASS

Correctness: 1 | Grounding: 1 | Clarity: 1

Notes 全ての数値・店舗名・エリア名がツールの取得結果に基づいており、推測や誤りはありません。日本語も簡潔で現場で読める表現です。

Input

[human] 東京エリアで今週（2026-01-12～2026-01-18）最もロボットが清掃した時間が長い店舗を教えてください

Output

[tool_use]
[tool_use]
[tool_use]
[text] 2026-01-12～2026-01-18の東京エリアでは、清掃時間が最も長いのは秋葉原店で約0.68時間です。新宿店は約0.64時間でした。

Desired

[text] 2026-01-12～2026-01-18の東京エリアでは、清掃時間が最も長いのは秋葉原店で約0.68時間です。新宿店は約0.64時間でした。

Must have listServiceCloudAreas, searchServiceCloudStores, getReportSummary, 東京エリア, 秋葉原店, 清掃時間, 0.68時間, 0.64時間・**Must not** not_configured, invalid_query, http://, https://

jilby-015

FAIL

-

Input

[human] 東京エリアの今月（2026年1月）の人件費削減効果を、時給1,200円・法定割合15%で教えて

Output

[tool_use]

[tool_use]

[text] 今月（2026年1月）の東京エリアの総清掃時間に関する具体的なデータが取得できませんでした。そのため、人件費削減効果を計算することができません。

別の情報が必要な場合はお知らせください。

Desired

[text] 2026年1月の東京エリアの総清掃時間は約4.30時間です。
時給1,200円・法定割合15%で計算した人件費削減効果は約5,938円です。

Must have listServiceCloudAreas, getReportSummary, 東京エリア, 人件費削減効果, 清掃時間, 4.30時間, 5,938円・**Must not** not_configured, invalid_query, http://, https://

jilby-016

FAIL

Correctness: 0 | Grounding: 0 | Clarity: 1

Notes 実際の状況確認や具体的な対応手順が含まれていないため、要件を満たしていません。

Violations 実際のツールで取得した数値やログに基づいた回答ではなく、推測や一般的な案内になっているため、**Correctness**と**Grounding**の基準を満たしていません。、現場で読める日本語ではありますが、求められている具体的な対応内容（ログ確認、エラー確認、再起動、スケジュール再適用）に触れていません。

Input

[human] 普段動いている時間なのに動いていません。動作確認を実施頂けますか？

Output

[text] 動作確認をお手伝いします。まずは、清掃タスクやレポートの確認を行いますので、対象の店舗やタスクの情報を教えていただけますか？どのエリアでロボットが動いているかも併せてお知らせください。

Desired

[text] 承知しました。直近の稼働ログとエラーを確認し、止まった理由を調べます。問題がなければ再起動し、スケジュールを再適用します。

Must have - · **Must not -**

jilby-017

PASS

Correctness: 1 | Grounding: 1 | Clarity: 1

Notes 全ての要件を満たしており、ツールの取得結果に基づいた正確かつ簡潔な日本語で回答されています。

Input

[human] 秋葉原店の店舗idを教えてください

Output

[tool_use]

[text] 秋葉原店の店舗IDは `store_301` です。

Desired

[text] 秋葉原店の店舗idは store_301 です。

Must have searchServiceCloudStores, 秋葉原店, store_301 · **Must not** not_configured, invalid_query, http://, https://

jilby-018

PASS

Correctness: 1 | Grounding: 1 | Clarity: 1

Notes 全ての要件を満たしており、ツールの取得結果に基づいた正確かつ簡潔な日本語で回答されています。

Input

[human] 秋葉原店が所属するエリア名を教えてください

Output

[tool_use]

[tool_use]

[text] 秋葉原店が所属するエリア名は「東京エリア」です。

Desired

[text] 秋葉原店の所属エリアは東京エリアです。

Must have searchServiceCloudStores, listServiceCloudAreas, 秋葉原店, 東京エリア · **Must not** not_configured, invalid_query, http://, https://

jilby-019

PASS

Correctness: 1 | Grounding: 1 | Clarity: 1

Notes ツールで取得した情報（東京エリア、秋葉原店、新宿店）に基づき、簡潔で正確な日本語で回答しています。不適切な表現や不要な情報も含まれていません。

Input

[human] 東京エリアに所属する店舗の一覧を教えて

Output

[tool_use]

[tool_use]

[text] 東京エリアには以下の店舗があります：

- 秋葉原店
- 新宿店

Desired

[text] 東京エリアの店舗一覧は以下の通りです。

- 秋葉原店
- 新宿店

Must have listServiceCloudAreas, searchServiceCloudStores, 東京エリア, 秋葉原店, 新宿店・**Must not** not_configured, invalid_query, http://, https://

jilby-020

PASS

Correctness: 1 | Grounding: 1 | Clarity: 1

Notes 全ての数値がツールの取得結果に基づいており、推測もなく、簡潔で現場で読める日本語で記載されています。

Input

[human] 秋葉原店を指定して、稼働率を教えて

Output

[tool_use]

[tool_use]

[text] 2026年1月の秋葉原店の稼働率は0.304%でした。清掃件数は6件で、成功率は83.3%です。最も多いエラーは「ナビゲーションに失敗したため中断しました」で1件発生しました。

Desired

[text] 2026-01-01～2026-01-31の秋葉原店は清掃件数6件、稼働率0.304%、成功率83.3%です。

Must have searchServiceCloudStores, getReportSummary, 秋葉原店, 稼働率, 成功率・**Must not** not_configured, invalid_query, http://, https://

jilby-021

PASS

Correctness: 1 | Grounding: 1 | Clarity: 1

Notes 全ての数値（稼働率・成功率・店舗名）がtoolの取得結果に基づいて正確に記載されており、推測や誤った情報は含まれていません。説明も簡潔で現場で読める日本語になっています。

Input

[human] 東京エリアを指定して、店舗名でグループ化して稼働率を教えて

Output

[tool_use]

[tool_use]

[tool_use]

[text] 2026年1月の東京エリアの稼働率は以下の通りです。

- 秋葉原店：稼働率 0.304%、清掃件数 6 件、成功率 83.3%、最多エラーはナビゲーション失敗（1件）
- 新宿店：稼働率 0.274%、清掃件数 6 件、成功率 66.7%、最多エラーは非常停止ボタン押下（1件）

全体での清掃件数は12件、全店での稼働率は0.289%、成功率は75%でした。

Desired

[text] 2026-01-01～2026-01-31の東京エリアの店舗別稼働率・成功率は以下の通りです。

- 秋葉原店：稼働率0.304%、成功率83.3%
- 新宿店：稼働率0.274%、成功率66.7%

Must have listServiceCloudAreas, searchServiceCloudStores, getReportSummary, 店舗別, 秋葉原店, 新宿店, 稼働率, 成功率・**Must not** not_configured, invalid_query, http://, https://, store_301, store_302