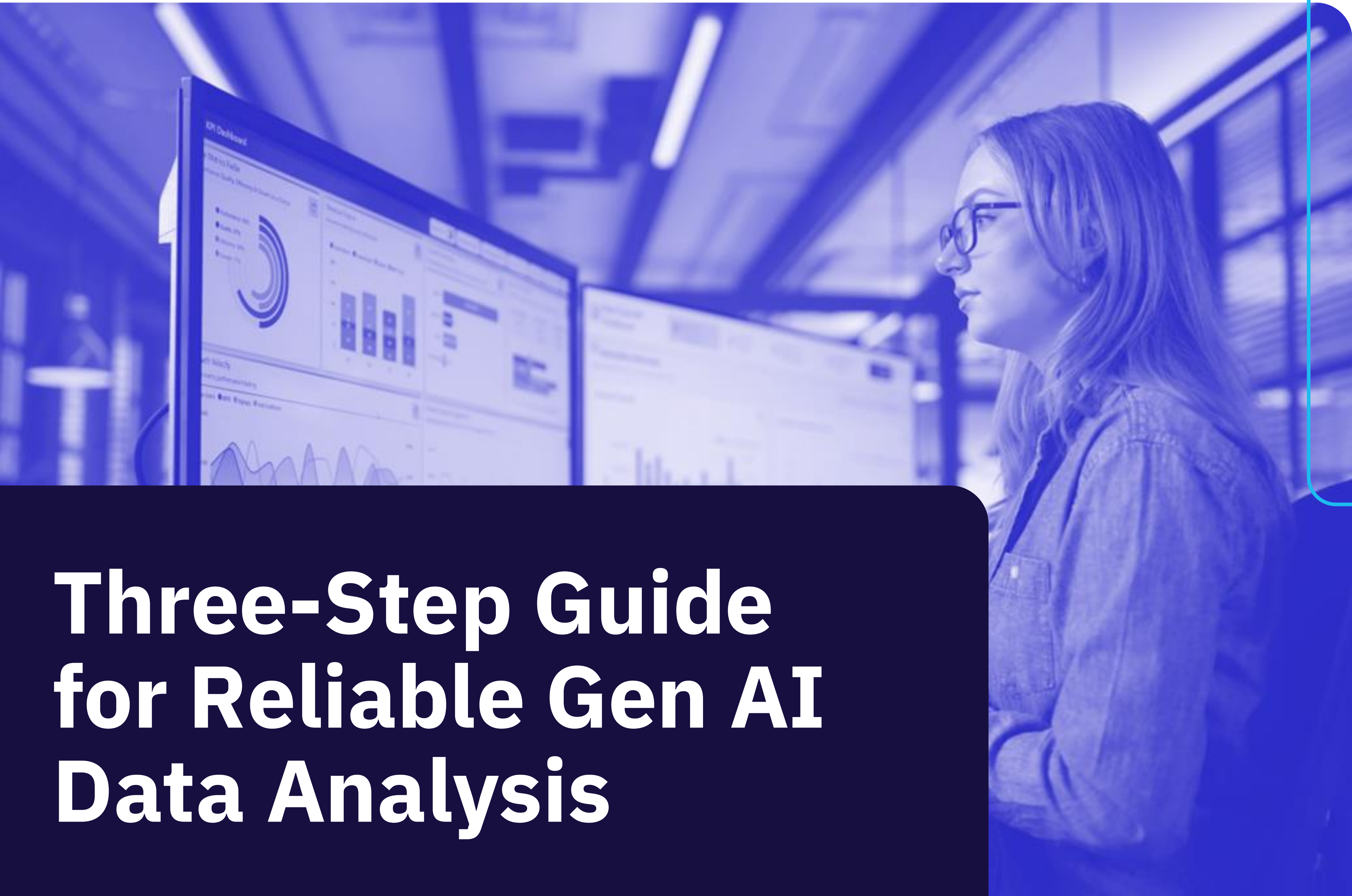

Three-Step Guide for Reliable Gen AI Data Analysis

// AIHR Resource

Introduction

These best practices enable you to analyze data accurately and robustly using LLMs by implementing safeguards. Using an iterative prompting approach ensures the LLM correctly accesses and interprets your data set, thereby mitigating risks such as hallucination and opaque reasoning.

Compliance check

Before uploading any company or sensitive data into an LLM, always ensure you are **fully compliant** with your organization's data security, privacy, and governance policies.

Step 1: Explore // Understanding, scoping, and validating the data

This phase focuses on establishing a foundation of trust with the LLM by clearly defining the analytical goal and verifying its understanding of the input data structure.

1 Establish hypotheses

Purpose

Use to explore your data and identify what you're looking for. Asking the LLM to explore the data without an underlying hypothesis risks type I errors (finding "significant" patterns when no true effect exists).

AI prompt

Before analyzing the data, I want to define what I am looking for. Help me articulate three to five clear, testable hypotheses based on the business question. For each hypothesis, specify which variables would need to be compared or calculated to confirm or refute it. Do not run any analysis yet.

2 Restate the data set structure

Purpose

Use to verify the LLM is correctly interpreting the data set structure. A full restatement prevents the model from hallucinating variables, misunderstanding data types, or missing important constraints.

AI prompt

Before performing any analysis, restate the data set as you understand it. List: i) total rows and columns, ii) each column name and data type, iii) min–max, mean, median for numeric fields, iv) unique categories for categorical fields, v) any missing values or anomalies.

Step 2: Test // Testing hypotheses and analyzing scenarios

This phase involves enforcing control mechanisms to ensure every calculation is validated before being presented.

3 Continuously enforce invariants during all calculations

Purpose

Use to ensure control totals (invariants) remain consistent and verifiable throughout every step of the analysis, providing an essential check against data corruption or faulty filtering.

AI prompt

Use the following control totals as invariants throughout the analysis:

[list invariants, e.g., Total Sales in USD, Total Customer Count]

Enforcement rule: During all calculations, ensure that any unfiltered aggregation of these invariants reconciles back to these control totals. If a total does not reconcile, stop, identify the source of the discrepancy, and correct it before continuing.

4 Check the logic and calculation plan

Purpose

Use to force the model to outline and justify its full computation path before running any analysis, preferably in an existing tool like Python code. This externalizes the internal logic and allows for human validation.

AI prompt

For each analysis you perform, first explain your calculation logic clearly: (i) list the formulas, filters, grouping logic, and denominators you are using, (ii) show the equivalent Python code for that calculation, and (iii) directly below it, provide the computed result. Do not separate these steps; they must appear as one integrated answer for each calculation.

5 Use redundant calculation paths

Purpose

Use to compute every key metric using at least two independent methods to catch silent errors and potential biases from a single method.

AI prompt

For each key metric, compute it using at least two independent calculation paths (e.g., direct aggregation versus group by; sum/count versus weighted mean).

For each path, show:

- The explanation of the logic
- The Python-style code
- The resulting value

After computing both values, compare them. If the values differ, stop the analysis, clearly flag the discrepancy, and explain possible reasons for the mismatch. Do not attempt to correct, adjust, or reconcile the values without explicit user guidance.

Step 3: Report // Translating data into validated insights

The final phase ensures the reported insights are both validated by multiple sources and sound based on domain knowledge.

6 Triangulate with multiple models

Purpose

Use to validate key findings by comparing them with results from an independent model or alternative analytical method.

AI prompt

Rerun the key metrics or conclusions using a different reasoning path or model (e.g., another LLM or an alternative analytical method). Present the new calculation logic, the Python-style code, and the resulting values. Compare the outputs from both analyses. If discrepancies appear, clearly flag them, explain potential reasons, and pause the analysis for user review rather than attempting to merge or reconcile the results.

7 Smell test the outcomes

Purpose

Use to ensure that results are plausible given the context, expected ranges, and domain norms. This is a critical check against nonsensical outputs.

AI prompt

Review all final results and evaluate their plausibility. For each key metric, state:

- The expected range based on the data set and domain knowledge
- Whether the computed value falls within that range
- Any results that appear inconsistent, extreme, or counterintuitive

Flag any suspicious or implausible outputs and provide possible explanations. Do not assume results are valid unless they pass these reasonableness checks.

Glossary: Essential Terminology

Core concepts

TERM	DEFINITION
Iterative prompting	A process of refining an LLM's output by providing continuous, structured feedback and constraints across multiple turns (prompts) to guide it toward a specific, accurate result, rather than relying on a single, complex initial prompt
Invariants (control totals)	Precalculated, unchanging summary metrics (e.g., total rows, total revenue, total users) from the source data set that are used as programmatic checks to ensure subsets, filters, and aggregations reconcile correctly throughout the analysis
Smell test	An informal, qualitative evaluation of data or analysis results to check for plausibility, common sense, and expected ranges based on domain knowledge. If a result "smells" wrong, it warrants deeper investigation.
Triangulate	The act of cross-validating a key finding or metric using two or more independent methods, data sources, or models. In LLM analysis, this often means comparing a result from the current LLM with one from another LLM or a traditional method.
Type I error	Also known as a false positive; the error of incorrectly concluding that an effect, relationship, or significant difference exists when, in reality, it does not (rejecting a true null hypothesis).
Type II error	Also known as a false negative; the error of incorrectly concluding that no significant effect, relationship, or difference exists when, in reality, it does (failing to reject a false null hypothesis).

Statistical tests

TERM	DEFINITION
***t*-statistic / *t*-test**	A t-test is a statistical test used to determine whether there is a statistically significant difference between the means of two groups, or between a sample mean and a known or hypothesized value. The t-statistic expresses the size of the difference relative to the variation in the data.
Chi-squared (χ^2)	A chi-squared test is a statistical test used with categorical data to assess whether observed frequencies differ significantly from expected frequencies. It is commonly used to test associations between categorical variables.

Cohen's *d*	A measure of effect size used to quantify the magnitude of the difference between two groups' means. It helps assess the practical significance beyond just the statistical significance.
Cramér's *V*	A measure of association (effect size) between two nominal (categorical) variables that have more than two categories. It is based on the Chi-squared statistic.

Variable types

VARIABLE TYPE	DEFINITION	EXAMPLES
Quantitative (numeric)	Variables representing meaningful numeric measurements where mathematical operations are valid.	Age, income, annual salary, years of service, overtime hours
Ratio scale	Highest level of measurement. Has all the properties of an interval scale plus a true, absolute zero point. Ratios are meaningful (e.g., 20 is twice 10). **True zero**: Zero means the absence of the quantity.	Annual salary (USD), number of open requisitions, employee age, years of experience
Interval scale	Has equal intervals between values but lacks a true zero point (zero is arbitrary). Ratios are not meaningful. **No true zero**: Division/multiplication is not valid.	Employee start date (e.g., 2024-01-15), performance review date

VARIABLE TYPE	DEFINITION	EXAMPLES
Categorical	Variables that represent groups, labels, or categories rather than meaningful numeric quantities.	Department, job role, marital status, location
Binary	A categorical variable with exactly two options or categories.	Attrition (yes/no), full-time status (yes/no), internal hire (yes/no), promotion (occurred/did not occur)
Nominal	A categorical variable where the categories have no intrinsic order or rank.	Marital status, department, region, gender
Ordinal	A categorical variable where the categories have a meaningful rank or order, but the distance between ranks is not equal or known.	Satisfaction rating (low, medium, high), performance rating (e.g., exceeds, meets, needs improvement), education level