

Prompt Testing and Evaluation Framework

Objective

To develop a structured, repeatable, and documented process for evaluating the performance and reliability of different Large Language Model (LLM) prompt versions for a specific task.

1. Task Definition and Success Criteria

Component	Description	Example (for a Summarization Task)
Task Objective	Precise, measurable statement of the desired outcome.	Generate a concise, neutral, 3-sentence summary of a provided news article.
Model & Parameters	The specific LLM and configuration (e.g., temperature, top-p).	Model: GPT-4o , Temperature: 0.3 (for determinism).
Key Performance Indicators (KPIs)	Formal metrics used to measure output quality.	1. Factual Accuracy (Score 1-5) 2. Neutrality/Tone (Score 1-5) 3. Compliance (Pass/Fail)
Success Threshold	The minimum acceptable performance for a prompt version.	Average Factual Accuracy score must be ≥ 4.0 , and the output must pass all Compliance checks.

2. Prompt Versioning and Structure

All prompt iterations must be treated as code, with structured components and version control.

A. Standard Prompt Structure (Template)

Component	Purpose	Example Directive

[ROLE/PERSONA]	Defines the AI's identity and perspective.	"You are a neutral, objective journalist."
[INSTRUCTION]	The core command defining the task.	"Analyze the following text and create a summary."
[CONTEXT/INPUT]	The variable data provided to the prompt.	[ARTICLE TEXT]
[FORMAT/CONSTRAINTS]	Specifies the required output structure and rules.	"Your summary must be exactly 3 sentences long and contain no subjective language. Output in Markdown format."

B. Version Control Log

Prompt ID/Version	Date	Change Log	Primary Author	Status
SUM-V1.0.0	YYYY-MM-DD	Initial baseline version.	J. Doe	Deprecated
SUM-V1.0.1	YYYY-MM-DD	Added a clear [ROLE] definition for better tone control.	J. Doe	Active
SUM-V1.0.2	YYYY-MM-DD	Refined sentence count constraint to use XML tags.	S. Smith	Testing

3.  Test Dataset Design

The evaluation dataset must be representative of production inputs.

Test Case ID	Input Type	Description	Expected Ground Truth Output	Purpose
TC-001	Standard	A well-written, straightforward article on a common topic.	A perfectly accurate, 3-sentence summary.	Baseline/Standard
TC-002	Edge Case	An article with highly technical jargon and inconsistent formatting.	A clear summary that simplifies the jargon.	Robustness
TC-003	Safety/Failure	A topic containing sensitive or highly controversial subject matter.	A summary that remains strictly neutral and avoids bias.	Guardrails
TC-004	Length Stress	An extremely long article (≥ 5,000 words).	A coherent summary within the length limit.	Token Limit/Context Handling

4. Evaluation Methods

Outputs are evaluated using a combination of automated and human/LLM-assisted review.

A. Automated/Compliance Metrics (Pass/Fail)

These are checks that can be run programmatically on the output.

Metric	Check	Compliance Requirement
Format	Is the output in the required format (e.g., JSON, Markdown)?	Must be TRUE

Length	Does the output comply with length limits?	\$\le 60\$ words OR exactly 3 sentences.
Punctuation	Does the output start with a capital and end with a period?	Must be TRUE
Redacted Terms	Does the output contain any prohibited words or phrases?	Must be FALSE

B. Human-in-the-Loop Evaluation (Scoring 1-5)

Subjective quality is assessed by human evaluators or by an "LLM-as-a-Judge" (a more advanced model).

KPI	Rating Description (1=Poor, 5=Excellent)
Factual Accuracy	5: All points are perfectly faithful to the source. 1: Contains significant hallucination or misrepresentation.
Neutrality/Tone	5: Strictly objective, no biased or emotional language. 1: Contains clear subjective opinion or bias.
Clarity/Coherence	5: Summary flows logically and is easy to understand. 1: Difficult to follow or poorly structured.

5. Results and Reporting

Field	Description
Test Run ID	Unique identifier for the full evaluation run.

Date of Run	When the test was executed.
Overall Performance	Prompt: SUM-V1.0.2 \$implies\$ Average Score: 4.7 (Best Performer)
Failure Analysis	Prompt: SUM-V1.0.1 failed TC-003 because it used a biased adjective ("shocking") when summarizing a sensitive topic.
Recommendation	SUM-V1.0.2 is approved for staging. Create a new Edge Case (TC-005) to test for adjective usage in all future runs.