1. Methodology: Prompt Formulation Strategy

A. Prompt Target (Data Sources)

1. Gemini-Only Projects: Exclusively targeting Gemini chat history and previous conversations.

2. Multi-Source Projects: Incorporating external personal data sources to elicit personalized responses, including:

- Gemini Chat History
- Gmail
- Google Photos
- Google Search History
- YouTube Watch History

B. Prompt Type

1. Explicit with source specified: Clearly dictating the exact data source to be used (e.g., *"From my Gmail..."*).

2. Explicit without source specified: Assuming the AI has personal context without stating where to look (e.g., *"When is my appointment?"*).

3. Implicit: Universal, generic queries designed to trigger personal context integration without containing any explicit personal references (e.g., *"How to prepare for an exam?"*).

4. Pre-created prompt: Standardized, system-provided queries (e.g., math problems or riddles) that typically do not expect a personalized output, yet the model may still integrate relevant personal context if applicable.

C. Prompt Category (Use Cases & Subareas)

Use Case (Broad Purpose)	Subareas (Specific Topic/Context)
Retrieval	Gemini App
Explanation	Edu, Current Affairs
Recommendations	Edu, Shopping, Travel/Local, Current Affairs, Media
Generation	Travel/Local, Edu, Shopping, Brainstorming, Planning
Self-Reflection	General, Specific, Attribution
Interpersonal / Relationships	Chit Chat, Advice, Simple pleasantries, Share sentiment, Perspective
Information Gathering	Fact & Simple search, Synthesize broad topics, Compile, Quizzes and tests
Translation	Native to Non-Native Language, Non-Native to Native Language
Capabilities	Attribution

2. Methodology: Side-by-Side (SxS) LLM Evaluation

A. Model Type (Evaluation Targets)

1. Test Model: The primary model version targeted for evaluation in its personalization capabilities, instruction adherence and other aspects in overall quality.

2. Base Model: The benchmark model utilized to establish a comparative performance baseline during the SxS analysis.

B. Conversation Type

1. Single-Turn: Evaluating immediate model response quality, direct instruction adherence, and pre-assigned task (translation of riddles, coding, math and other scientific items).

2. Multi-Turn: Assessing conversational memory, context awareness and retention, and adaptability over extended, multi-step dialogue interactions.

C. Critical Inspection Areas

1. Sian Profile in Debug Info: Inspecting the extracted data to verify the accuracy of the retrieved personal context and ensure no forced personalization.

2. Source Profile in Debug Info: Verifying model routing decisions regarding the source's steps execution.

3. Grounded Assessment in Response: Cross-referencing AI outputs against existed personal data *(see Page 3 for deeper aspects in the evaluation)*.

D. Audit Trail Type

1. Debug Info Extraction: Systematic retrieval of backend .txt logs for every conversational turn to verify active Model IDs.

2. Full HTML Export: Preserving immutable, complete webpage records of the entire conversation thread for Quality Assurance (QA) verification.

Key Technical Protocols applied in this workflow:

- **Backend Auditing:** Verification of Model IDs through Debug Info.
- **Personalization Verification:** Deep-dive into Sian Profile to confirm accurate personal data and seamless integration.

3. AI Responses Evaluation Matrix

A. Objective Rating Type

1. Personalization Presence: Objectively verifying if the model successfully utilized personal data or style preferences from out-of-current-session historical context.

Sub-components:

- Personalized with User Personal Data
- Personalized with Style Preference

2. Personalization Groundedness: Verifying that the personalized elements generated are strictly anchored in factual user data without hallucinations.

3. Personalization Adaptability: Assessing the model's ability to seamlessly integrate context shifts and user corrections and constraints within the ongoing prompt.

B. Subjective Rating Type

1. Personalization Appropriateness: Evaluating whether the use of personal context is socially appropriate, natural, and non-intrusive for the given task.

2. Personalization Helpfulness: Assessing if the personalized response genuinely adds value, efficiency, or utility to the user's specific request.

3. Total Personalization Quality: A composite evaluation combining Appropriateness and Helpfulness to determine the overall effectiveness and impact of the personalization.

C. Overall Quality

An assessment of the conversation as a whole, considering personalization alongside the following general dimensions:

- **Safety & Harmlessness:** Ensuring the AI strictly adheres to safety guidelines, producing responses free from toxicity, bias, or harmful content.
- **Instruction & Constraint Adherence:** Whether the AI correctly followed all instructions, including strict adherence to negative constraints, formatting rules, and length limits.
- **Collaborativity:** How well the AI collaborated and helped move the conversation forward with suggestions and possible next steps.
- **Writing Style/Tone:** Whether the responses were well written, utilizing high-quality, engaging, and digestible conversational prose, with suitable communication format and style (formal for scientific report and casual for conversational with personalized style preference if any).
- **Contextual Awareness:** How well the AI remembered and built on information from previous turns in the conversation as the interaction progressed.
- **Content Relevance:** How relevant the content provided by the AI was towards accomplishing the specific goal.
- **Content Completeness:** How complete the provided content was (i.e., having enough detail so that no follow-up was needed).
- **Truthfulness:** How truthful the AI's responses were, based on real-world knowledge.

Evaluation Core Principle:

A response may be highly personalized but fail in overall quality if it hallucinates facts, provides generic filler, or violates critical safety guidelines. High quality personalization cannot rescue a fundamentally flawed response.

4. Quality Assurance: Rationale & Justification Writing

A. Core Principles of Evaluation Rationales

1. Comparative Analysis: Rationales must explicitly compare the Test Model and Base Model against each other, highlighting the delta in their performance rather than evaluating them in isolation.

2. Evidence-Based Claims: Every rating must be substantiated with direct quotes, specific data points, or concrete examples extracted directly from the models' outputs.

3. Metric Alignment: Justifications must directly address the specific dimensions (e.g., Truthfulness, Personalization Helpfulness) that drove the final Overall Quality score.

B. The Anatomy of a Standard Justification

A high-quality rationale is systematically structured into four key components:

- **The Verdict:** A clear, upfront statement declaring which model performed better and the degree of difference (e.g., *"Model A is slightly better than Model B"*).
- **The Strengths:** A detailed breakdown of what the winning model did successfully, specifically regarding instruction following and content relevance.
- **The Weaknesses:** A critical analysis of where the losing model fell short (e.g., identifying hallucinations, generic filler, or formatting failures).
- **Personalization Deep-Dive:** An explicit assessment of how well personal data was integrated, including its naturalness and adherence to safety constraints.

C. Real-World Comparative Rationale Excerpt (Example)

"[Turn 1] Both Model A and Model B responses were helpful well-structured with numbers and bullets, increasing readability and quicker to understand for the reader. Both of them also provided complete factual 6 main points of the main argument from the author of the article which was good in truthfulness and instruction following. Both of them placed the main argument at the end of their responses instead of at the beginning, which was a minor issue in following instruction as the prompt actually asked for the main argument and therefore, it should be ideally placed at the beginning. Both model showed good writing style as it was sufficiently formal yet easy-to-understand for an article review. However, Model A made much worse issue of providing wrong information (if it was not going to be referred as fake) about the republishing of the article under another title of 'Homeschooling: A Growing Option in American Education' WHILE Model B made less serious issue when using misleading language as if the author was also the owner of Heritage Foundation. So, I rated Model B as better for Quality Comparison SxS."

Quality Assurance Mindset:

A rating without a robust rationale is merely an opinion. A comprehensive, evidence-driven rationale transforms that rating into actionable engineering feedback.