SER 517 - Sprint 1

Task #25 - Design Template Structure for Auto - Response Generation

User Story #24 - Develop Prompt Templates

Epic #1 - Determine Optimal Prompt Strategies for Generating Auto - Responses from User Availability Data

Author: Prerana Kumsi

ASU ID: 1233380995

1. Objective

The purpose of this task is to design a standardized prompt template structure for generating contextual auto - responses. The template is grounded in empirical findings from previous Sprint 1 tasks that analyzed:

- Which contextual elements are essential
- Which elements significantly improve response quality
- Which elements produce diminishing returns
- How privacy and hallucination risks interact with context inclusion

The final template must:

- Maximize response quality and clarity
- Minimize unnecessary context and privacy exposure
- Reduce hallucination and over - specification
- Support tone adaptation based on sender relationship
- Remain modular and extensible for Stage 2 API integration

2. Design Principles Derived from Previous Analysis

2.1 Activity Is the Core Driver

Current activity is the single most important contextual element. Without activity, responses become vague and unjustified.

Example of degradation:

- With activity → "He is in a scheduled meeting."
- Without activity → "He may not be available."

Therefore, Current Activity must always be included.

2.2 Optimal Context Set = 5 High - Impact Elements

Combination testing revealed that the following five elements produce near - maximum response quality:

1. Current Activity
2. Current Time
3. Sender Relationship
4. Message Urgency

5. Expected Response Time / Duration

Adding additional environmental or device - level context (noise, light, battery, lock state) provided minimal to no improvement.

This establishes a clear quality inflection point at 5 elements.

2.3 Diminishing Returns Beyond Tier 2 Context

Adding low - impact elements results in:

- Increased token cost
- Increased privacy exposure
- Risk of over - specification
- No meaningful improvement in output quality

Therefore, the template must prioritize:

- Tier 1 (Critical) elements
- Tier 2 (High - impact) elements

Lower - tier elements should only be included if explicitly required.

2.4 Privacy as a Cross - Cutting Filter

Before assembling the prompt:

- Sensitive locations must be redacted
- Confidential meeting details must be suppressed
- User - defined sharing rules must override defaults
- Data minimization settings must limit context inclusion

Privacy is applied before the prompt reaches the LLM, not after generation.

3. Template Design Requirements

The prompt template must:

- Use structured context (not raw sensor values)
- Provide high - level inferred activity (e.g., "in a meeting" instead of "noise: 72dB")
- Avoid passing raw GPS coordinates
- Avoid exposing device - level micro - context
- Adjust tone based on sender relationship
- Explicitly instruct the LLM not to fabricate details
- Speak about the user in third person

4. Final Prompt Template Structure

4.1 System Instruction Layer (Constant)

You are an auto - response generation system.

Generate a polite, context - aware auto - response.
- *Speak about the user in third person.*
- *Do NOT imitate the user directly.*

- *Do NOT fabricate details.*
- *Use only the supplied context.*
- *Adjust tone based on sender relationship.*
- *Set realistic expectations for response timing.*
- *Do not add unnecessary details.*

This instruction:

- Reduces hallucination
- Prevents over - specification
- Enforces consistent voice

4.2 Structured Context Block

Required (Always Include)

User Name:
Current Activity:
Current Time:

These elements ensure:

- Clear reason for unavailability
- Temporal grounding
- Meaningful justification

Recommended (Include When Available)

Sender Relationship:
Message Urgency:
Expected Response Time / Duration:

These elements:

- Calibrate tone (manager vs friend)
- Clarify urgency handling
- Set expectation for follow - up

Optional (Privacy - Dependent)

Semantic Location Label:
Day of Week:
User Tone Preference:

These may:

- Add personalization
- Improve natural phrasing
- Enhance contextual nuance

But should only be included if privacy settings permit.

Explicitly Excluded by Default

- Noise level

- Light level
- Battery level
- Device lock state
- Network connectivity
- Raw GPS coordinates
- Crowd density
- App currently active

These:

- Provide negligible response improvement
- Increase privacy risk
- Add unnecessary tokens

5. Template Variants Based on Context Sensitivity

5.1 Minimal (Privacy - Optimized)

Include:

- Current Activity
- Expected Response Time

Use case:

- High privacy sensitivity
- Minimal disclosure preference

Produces:

- Functional but neutral response

5.2 Standard (Recommended Default)

Include:

- Current Activity
- Current Time
- Sender Relationship
- Message Urgency
- Expected Response Time

Produces:

- Excellent quality
- Natural tone adaptation
- Clear expectation setting
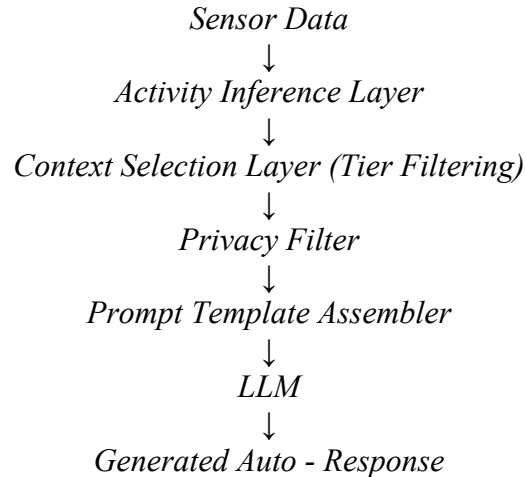- Strong privacy - to - quality balance

5.3 Rich (Enhanced Context)

Include:

- Tier 1 + Tier 2 + selected Tier 3 elements

Use case:

- User explicitly allows personalization
- Socially adaptive scenarios

6. Prompt Assembly Flow (Architecture)

Sensor Data

↓

Activity Inference Layer

↓

Context Selection Layer (Tier Filtering)

↓

Privacy Filter

↓

Prompt Template Assembler

↓

LLM

↓

Generated Auto - Response

Key Principle:

High - level inferred context is passed to the LLM - not raw sensor values.

7. Example Instantiated Template

Input Context

User Name: Satyam
Current Activity: In a scheduled team meeting
Current Time: 2:15 PM Tuesday
Sender Relationship: Manager
Message Urgency: Medium
Expected Response Time: ~30 minutes

Generated Output

Satyam is currently in a scheduled team meeting and is unable to respond at this time. He is expected to be available in approximately 30 minutes and will review your message promptly once the meeting concludes.

8. Alignment with Key Research Questions

Research Question	Template Alignment
How should user context be represented to the LLM?	Template uses structured, high - level inferred activities rather than raw sensor values.
How much context is necessary vs sufficient?	Template establishes 5 optimal elements as the quality inflection point, with clear tiers for progressive enrichment.

Does increasing specificity improve clarity or increase privacy risk?	Template includes privacy filter layer and explicitly excludes low - value, high - risk elements like GPS coordinates and device states.
How does context selection affect hallucination?	System instructions explicitly prohibit fabrication and limit LLM to supplied context only.

9. Empirical Grounding from Previous Tasks

This template structure is directly informed by findings from:

Task	Key Finding Applied to Template
Task #17 (Prerana Kumsi)	Comprehensive contextual element taxonomy provided foundation for element selection.
Task #19 (Aniket Patil)	Individual element removal tests established activity as critical; device state and environmental micro - context as unnecessary.
Task #20 (Aniket Patil)	Combination testing identified 5 - element optimal set and demonstrated diminishing returns beyond it.
Task #21 (Satyam Shekhar)	Context impact consolidation validated critical vs high vs minimal impact classification.
Task #22 (Satyam Shekhar)	Priority ranking established tier structure and privacy - as - filter approach.

10. Stage 2 API Integration Considerations

The template structure is designed with API integration in mind:

10.1 Modular Context Assembly
- Context blocks can be dynamically assembled based on availability
- Template supports graceful degradation when elements are missing
- Each tier operates independently

10.2 Privacy Configuration
- User privacy settings can map to template variants (Minimal/Standard/Rich)
- Privacy filter operates before prompt assembly
- Sensitive data redaction occurs at infrastructure level

10.3 Extensibility
- New contextual elements can be added to appropriate tiers
- System instructions remain constant across updates
- Template supports A/B testing of context combinations

10.4 Performance Optimization

- Minimal variant reduces token count for cost efficiency
- Standard variant balances quality and token cost
- Caching system instructions reduces repeated tokens

11. Validation Strategy for Stage 2

The template should be validated through:

- **Quality metrics:** Naturalness, clarity, appropriateness (rated by human evaluators)
- **Privacy metrics:** Information disclosure, sensitivity violations
- **Hallucination metrics:** Fabricated details, over - specification instances
- **Performance metrics:** Token count, latency, cost per response
- **User acceptance:** Would users send this response? Does it match their communication style?

12. Conclusion

The designed prompt template structure is grounded in empirical testing and context impact analysis from Tasks #17 - #22. It represents a data - driven approach to context selection that:

- Prioritizes activity as the primary driver of response clarity
- Uses five optimal high - impact elements for the standard configuration
- Avoids low - value environmental and device - level data
- Minimizes privacy exposure through tiered filtering
- Reduces hallucination risk through explicit system instructions
- Supports modular extensibility for Stage 2 API implementation

This template provides a scalable and privacy - aware foundation that balances response quality with data minimization. The three - tier variant system (Minimal, Standard, Rich) enables flexible deployment based on user preferences and privacy requirements.

The standard 5 - element configuration represents the optimal balance point identified through systematic combination testing in Task #20, achieving excellent response quality while excluding 44% of available contextual elements that provided no meaningful improvement.

13. References

- Task #17 - List All Possible Contextual Elements (Prerana Kumsi)
- Task #19 - Testing Prompt Robustness by Removing Individual Contextual Elements (Aniket Patil)
- Task #20 - Testing Prompt Robustness with Combinations of Contextual Elements (Aniket Patil)
- Task #21 - Context Impact on Response Quality (Satyam Shekhar)
- Task #22 - Priority Ranking of Contextual Elements (Satyam Shekhar)
- Task #23 - Context Selection Guidelines (Satyam Shekhar)