📋 Draft Research Plan

Prompt Engineering Virtual Internship – Week 1

## **Research Plan for Prompt Engineering Virtual Internship - Weeks 2-4**

**Research Objectives**

The primary objectives for the next phases of the internship for our team are:

1. **Evaluating different AI tool performances:**

We will examine how OpenAI, Cohere, Claude, Microsoft Azure AI, Hugging Face, and Google Gemini perform in specific prompt engineering tasks. This will help us find the best use cases.

1. **Optimize prompt design:**

We will create and improve prompts to enhance output quality. We will focus on clarity, specificity, and context for each tool.

1. **Explore customization and integration:**

We will investigate the customization options and integration possibilities of each tool for learning programs and business applications.

1. **Identify best practices:**

We will set up best practices for prompt engineering that cater to different user skill levels, including beginner, intermediate, and advanced, along with various use cases.

**Focus Areas**

Based on the Comparative Analysis Report, the following criterions of prompt engineering will be researched upon to identify the most feasible AI tool:

1. **Customization**:
   1. Investigate to what extent customized models in Hugging Face can be used, as well as developer-level customization in Cohere and Azure AI.
   2. Explore prompt engineering methods to create the feel of customization in tools that are not very flexible, such as OpenAI, Claude, and Gemini.
2. **Tool Integration**:
   1. Test how well these tools integrate with external systems. Look at Microsoft services for Azure AI, Google apps for Gemini, and APIs for OpenAI and Hugging Face.
   2. Evaluate how this integration affects usability and output quality in both business and learning settings (learning experience).
3. **Output Quality and Consistency**:
   1. Check the quality of the output for specific tasks, such as creative writing, summarization, and classification, across different tools.
   2. Assess how consistent the responses are when prompts vary in structure and complexity.
4. **Multimodal Capabilities**:
   1. Explore OpenAI’s support for different types of content, including text, images, and audio, to see its range of applications.
   2. Compare this with other tools that do not have multimodal features, like Cohere, Claude, and Gemini, to identify the trade-offs.
5. **Ethical and Safe Responses**:
   1. Investigate Claude’s method for generating ethical content and avoiding bias.
   2. Compare this with other tools for providing safe and responsible AI outputs in educational programs.

**A Summarized table for the above focus areas and Research Objectives:**

| **Research Question** | **Priority (1-5)** | **Justification** | **Type of Data** | **Est. Date of Completion** | **Client Needs** |
| --- | --- | --- | --- | --- | --- |
| **Customization** | | | | | |
| How can prompt engineering simulate advanced customization in OpenAI and Google Gemini for creative writing tasks? | 5 | OpenAI and Gemini lack deep customization, but prompt engineering can enhance output specificity for educational content as well as content as per the user’s needs. | Qualitative (text output quality), Quantitative (prompt iteration count) | July 4, 2025 | Flexible, high-quality content for learning programs. |
| What are the optimal fine-tuning strategies for Hugging Face models in document classification tasks? | 4 | Hugging Face’s open-source models allow extensive customization, critical for enterprise NLP accuracy. | Quantitative (classification accuracy, training time), Qualitative (ease of fine-tuning) | July 7, 2025 | Custom AI solutions for enterprise document management. |
| How does Cohere’s developer-level customization compare to Azure AI for semantic search applications? | 4 | Both tools offer enterprise-focused customization, but their effectiveness in search tasks requires comparison. | Quantitative (search precision, recall), Qualitative (setup complexity) | July 9, 2025 | Efficient search solutions for enterprise workflows. |
| **Tool Integration** | | | | | |
| How seamlessly can OpenAI integrate with external APIs for chatbot development in learning environments? | 5 | OpenAI’s broad API support is ideal for chatbots in educational settings, but integration ease needs validation. | Qualitative (integration success, user feedback), Quantitative (API response time) | July 6, 2025 | Scalable chatbot solutions for education. |
| How effective is Azure AI’s integration with Microsoft Teams for customer support applications? | 4 | Azure AI’s Microsoft ecosystem integration is key for business use cases like customer support. | Qualitative (user experience, integration errors), Quantitative (sentiment analysis accuracy) | July 10, 2025 | Seamless business application integration. |
| Can Google Gemini integrate with Google Workspace for quick summarization tasks in educational settings? | 3 | Gemini’s integration with Google apps could streamline content creation for non-technical users. | Qualitative (integration usability), Quantitative (summarization speed) | July 11, 2025 | User-friendly tools for educational content creation. |
| **Output Quality** | | | | | |
| How does Claude’s output quality compare to OpenAI for structured writing tasks like report generation? | 5 | Claude’s strength in structured writing could outperform OpenAI for polished educational content. | Qualitative (text coherence, structure), Quantitative (editing time required) | July 5, 2025 | High-quality, polished content for reports. |
| What factors influence Cohere’s output consistency in document classification tasks? | 4 | Consistent outputs are critical for enterprise NLP reliability, especially in classification. | Quantitative (classification consistency, error rate), Qualitative (output stability) | July 8, 2025 | Reliable outputs for enterprise applications. |
| How does Gemini’s output quality perform for quick summarization compared to Claude? | 3 | Gemini’s beginner-friendly interface may offer faster summarization for casual users. | Qualitative (summary clarity), Quantitative (summarization speed) | July 10, 2025 | Fast, clear summaries for educational use. |
| **Multimodal Capabilities** | | | | | |
| How effective is OpenAI’s multimodal support for combining text and image inputs in creative tasks? | 5 | Multimodal capabilities expand creative applications, critical for engaging learning content. | Qualitative (output creativity, relevance), Quantitative (task completion rate) | July 7, 2025 | Engaging, versatile content for education. |
| What are the limitations of non-multimodal tools (Cohere, Claude) in tasks requiring image context? | 4 | Understanding single-modality tool limitations helps define their scope in learning programs. | Qualitative (task feasibility analysis), Quantitative (task failure rate) | July 9, 2025 | Clear scope for tool selection in education. |
| Can prompt engineering compensate for the lack of multimodal support in Gemini for creative tasks? | 3 | Prompt engineering may mitigate Gemini’s lack of multimodal features for specific use cases. | Qualitative (output quality), Quantitative (prompt complexity) | July 11, 2025 | Accessible creative tools for non-technical users. |
| **Ethical Responses** | | | | | |
| How does Claude’s ethical response framework handle sensitive topics compared to OpenAI? | 5 | Ethical responses are critical for safe content in educational settings. | Qualitative (bias analysis, safety rating), Quantitative (response rejection rate) | July 6, 2025 | Safe, unbiased content for learning programs. |
| How do Hugging Face models perform in terms of bias mitigation when fine-tuned for content generation? | 4 | Fine-tuning Hugging Face models could reduce bias, but effectiveness needs testing. | Qualitative (bias presence), Quantitative (bias detection metrics) | July 8, 2025 | Responsible for AI for educational and enterprise use. |
| How does Azure AI’s sentiment analysis ensure ethical responses in customer-facing applications? | 3 | Azure AI’s emotion detection must align with ethical standards for business applications. | Qualitative (response appropriateness), Quantitative (sentiment accuracy) | July 10, 2025 | Ethical, customer-friendly responses for businesses. |

**Methodology**

The research will follow a structured, iterative approach to test tools, refine prompts, and analyze results. The steps are:

1. **Task Definition and Use Case Selection**:
   1. Select three main use cases from the report: chatbots/virtual assistants, content generation/rewriting, and document classification/semantic search.
   2. Define specific tasks for each use case (e.g., generating a 200-word blog post for content generation, classifying customer feedback for semantic search).
2. **Prompt Design and Testing**:
   1. Create a standard set of prompts for each tool, with varying levels of complexity (e.g., simple, intermediate, advanced).
   2. Test the prompts across all tools to assess output quality, relevance, and consistency.
   3. Refine the prompts based on early results to improve performance (e.g., adjusting tone, specificity, or context).
3. **Tool-Specific Customization**:
   1. For Hugging Face, fine-tune a pre-trained model using a small dataset related to the internship learning program.
   2. For Cohere and Azure AI, explore developer-level customization options (e.g., embeddings for Cohere, sentiment analysis for Azure AI).
   3. Simulate customization in OpenAI, Claude, and Gemini using advanced prompt engineering methods.
4. **Integration Testing**:
   1. Test API-based integrations for OpenAI and Hugging Face in a Python environment to mimic enterprise workflows.
   2. Evaluate Azure AI’s integration with Microsoft Teams or Power Apps for business use cases.
   3. Check Gemini’s compatibility with Google Workspace for quick content creation tasks.
5. **Performance Evaluation**:
   1. Measure output quality using metrics like relevance, coherence, and task completion rate.
   2. Gather qualitative feedback from team members to evaluate usability and beginner-friendliness.
   3. Compare tools based on speed, cost (for non-free tiers), and scalability for enterprise or learning applications.
6. **Data Collection and Analysis**:
   1. Collect outputs from each tool for the defined tasks and store them in an organized format (e.g., CSV or database).
   2. Analyze results to find patterns, such as which tools perform well on specific tasks or need less prompt refinement.
   3. Document limitations (e.g., hallucination in OpenAI, lack of emotion detection in Claude), and their effects on use cases.
7. **Reporting and Recommendations**:
   1. Put together findings into a mid-internship report that summarizes tool performance, prompt improvement strategies, and integration results.
   2. Offer practical recommendations for choosing tools based on use case, technical skills, and resource availability.
   3. Share best practices for prompt engineering to assist future interns or enterprise users.

**Evaluation Metrics**

**Objective:**

To systematically assess the performance, usability, and integration potential of selected AI tools (OpenAI, Hugging Face, Cohere, Claude, Azure AI, Google Gemini) across various prompt engineering tasks.

1. **Output Quality**

| **Sub-metrics** | **Definition** | **Measurement** |
| --- | --- | --- |
| **Relevance** | How closely the response matches the task or prompt intent. | Human rating (1–5); semantic similarity to reference. |
| **Coherence & Fluency** | Whether the text flows logically and is grammatically correct. | Human rating + Grammarly or similar NLP tools. |
| **Completeness** | Whether all required elements of the task are present. | Checklist evaluation or % of elements matched. |
| **Task Appropriateness** | Suitability of output for its intended domain (e.g., educational, enterprise). | Expert judgment + scenario testing. |

1. **Output Consistency**

Evaluates how stable the model’s responses are under varying prompt conditions.

| **Sub-metrics** | **Definition** | **Measurement** |
| --- | --- | --- |
| **Stability across reruns** | Output variability when same prompt is run multiple times. | % of matching key elements; edit distance. |
| **Structure sensitivity** | Impact of slight prompt rewording on output quality. | Deviation score from original; consistency index. |
| **Response diversity** | For creative tasks, measures how varied the outputs can be. | Diversity index (e.g., lexical variation ratio). |

1. **Customization Potential**

Assesses how well each tool supports or simulates customization.

| **Sub-metrics** | **Definition** | **Measurement** |
| --- | --- | --- |
| **Model fine-tuning** | Ability to train or adapt models to a specific task or dataset. | Training time, accuracy gain, and documentation clarity. |
| **Prompt flexibility** | Degree of control through prompt design alone. | # of iterations needed to reach optimal performance. |
| **API/SDK support** | Availability of developer tools for customization. | API documentation completeness; ease-of-use score. |

1. **Integration Capability**

Measures how easily tools integrate with other systems for practical workflows.

| **Sub-metrics** | **Definition** | **Measurement** |
| --- | --- | --- |
| **Tool ecosystem fit** | Native integrations (e.g., Azure with MS Teams; Gemini with Google Docs). | Integration success rate; setup time. |
| **External API support** | Ease of using REST APIs in external applications. | API response time; authentication/connection stability. |
| **Workflow compatibility** | How well outputs serve business/learning pipelines. | Time saved in workflow; user ratings from pilot tests. |

1. **Multimodal Support**

Evaluates performance in tasks that involve combining text, image, or other data types.

| **Sub-metrics** | **Definition** | **Measurement** |
| --- | --- | --- |
| **Content Fusion Quality** | How well a tool synthesizes inputs from multiple modalities (e.g., image+text). | Task success rate; user-rated relevance. |
| **Fallback Prompting** | How effectively text-only tools simulate multimodal context. | Prompt complexity score; output accuracy rating. |

1. **Ethical & Safe Responses**

Checks how responsibly tools handle sensitive topics or biased input.

| **Sub-metrics** | **Definition** | **Measurement** |
| --- | --- | --- |
| **Bias Mitigation** | Absence of gender, racial, or political bias. | Bias audit (toxicity/BERTScore); human feedback. |
| **Sensitive Content Filtering** | Appropriateness of output for minors or general audiences. | Rejection rate; compliance with safety frameworks. |
| **Transparency** | Whether the tool gives rationale or disclaimers for sensitive content. | Binary (yes/no); ethical compliance checklist. |

1. **Performance & Cost Efficiency**

Compares tools based on speed, affordability, and scalability.

| **Sub-metrics** | **Definition** | **Measurement** |
| --- | --- | --- |
| **Latency** | Time to first response. | Milliseconds or seconds per request. |
| **Throughput** | Number of prompts handled per minute/hour. | Prompts per unit time. |
| **Cost per Output** | Price (if applicable) per generated or evaluated prompt. | USD/token or USD/hour (from pricing tiers). |
| **Scalability** | Suitability for high-volume or enterprise deployment. | API rate limits, concurrent request limits. |

**Scoring Approach:**

Each metric will be scored on a **1–5 scale** or normalized where appropriate. Weights will be applied depending on **task type**:

| **Task Type** | **High-Priority Metrics** |
| --- | --- |
| Content Generation | Output Quality, Ethical Responses, Multimodal Support |
| Semantic Search | Customization, Output Consistency, Tool Integration |
| Chatbots | Integration Capability, Response Coherence, Latency |
| Classification | Output Accuracy, Consistency, Fine-tuning Support |

**Timeline**

* **Week 2**: Task definition, initial prompt design, and baseline testing of all tools.
* **Week 3**: Customization experiments, integration testing, and iterative prompt refinement.
* **Week 4**: Performance evaluation, data analysis, and drafting of the final week's deliverables.

**Expected Outcomes**

* A detailed comparison of tool performance across selected use cases. (Comparative Analysis Report, inserted previously) and to be refined and worked upon.
* A set of optimized prompts tailored to each tool and task.
* Recommendations for tool selection and prompt engineering best practices for learning programs and enterprise environments.
* Insights into customization and integration challenges to inform future research.