

Generative AI Lab Project Plan

BMW Group: Agent Evaluation and Automated Prompt Optimization Pipeline

Barath Velmurugan, Bernardo Chalita, Erwin Deng, Ivy Xu

1. Project Overview

We are building an automated pipeline that evaluates AI agent performance and continuously optimizes their system prompts based on feedback, without manual intervention or model retraining. The goal is to enable BMW's deployed agents to adapt to evolving tasks and data across business units by learning what "good output" looks like and updating prompts accordingly.

2. Scope

This is an experimental (research) project, not a production system (as defined by the BMW team). We will focus on one or two document/data types (e.g. car reports) and build the evaluation and prompt optimization pipeline around them. The emphasis is on the pipeline's ability to generalize and evolve prompts across different document types, not on achieving specific metrics (e.g., high accuracy). Reinforcement learning and model retraining are explicitly out of scope, as the objective is to achieve comparable benefits through prompt optimization alone.

3. Milestones

- *Now to March 7:* Literature review (Barath recommended paper for example: <https://arxiv.org/abs/2507.19457>), define evaluation metrics, set up dev environment (Openrouter, Ollama, Qwen, **Colab**), identify 1-2 data types (i.e., documents) to work with
- *March 7 to March 18:* Baseline agent implementation with initial system prompt; first evaluation framework running against ground truth; benchmark baseline scores
- *March 18 to March 29:* DS Trek and vacation, limited availability, plan for this buffer
- *March 29 to April 7 (Midpoint):* Automated prompt optimization loop running end-to-end; pre vs. post optimization benchmarks showing measurable improvement; clear narrative on methodology
- *April 7 to April 28:* Test pipeline generalizability on second document/data type; refine evaluation framework; address stability and edge cases
- *April 28 to May 13 (Final):* Final benchmarks, documentation, and presentation; clean demo of full pipeline

4. Timeline

By the midpoint presentation (April 2 or 7), we want a working end-to-end pipeline: an agent with a baseline system prompt, an evaluation framework comparing outputs to ground truth, and at least one

round of automated prompt optimization with measurable improvement in eval scores. The final presentation on May 13 should show generalizability across data types and a polished demo of the full self-evolving pipeline.

5. Team & Responsibilities

Specific role assignments are still being scoped, but work can be parallelized across a few tracks: one subgroup focuses on the evaluation framework and metrics design, another on the prompt optimization loop and LLM integration, and optionally a third on data preparation and benchmarking. Our mentor Fiona (former MBAn, did her capstone project with BMW Group) will provide guidance on structuring progress and keeping scope realistic. BMW POCs Jason (technical) and Dave (strategy) will advise on domain requirements and data access.

6. Tools to use

We plan to use Ollama with small open-source models (e.g. LLaMA 3.2, vision models) for local inference and Google Colab for compute/coding. Pipeline development will be in Python, and we will explore prompt optimization frameworks such as GEPA through DSPy (however, we will expand more as we conduct our literature review). No BMW proprietary infrastructure is assumed for now. Data will be sample documents such as car reports, either provided or approved by BMW.

7. Questions to ask your mentors and host POC

Once we get the data, we may want to understand the dataset in full so we know exactly what is included in the pre/post-optimization benchmarks. BMW has agreed to provide context with this data (from our initial call). We will also continue refining our approach based on our own research and literature review. New questions will likely continue to come up as we align our ideas with the company's business goals. Since they mentioned that they care less about metrics (such as accuracy), and more about the overall framework, we want to keep clarifying how evaluation should be done as it is dynamic and critical for this research-based project (to validate the framework).

8. Additional Information

- We met with the host's primary contact and our mentor by February 20.
- We will receive data access this week.
- The team has created a GitHub repository and Colab workspace for collaboration.