Isophore: LLM Metrology from Precogrecog

Useful measurement is at the heart of robust and reliable AI systems

Prepared by: precogrecog.com
Date: February 2026

Executive Summary: The way that AI application performance is normally assessed is deeply flawed. Isophore is a new software tool that can be used to create useful and informative measurements of AI performance in a specific application.

- How robust is your prompt; is it likely to fail in prod?
- How likely it is that there's a better prompt?
- How much effort you should put into further prompt optimisation?

1 Introduction

These implementations are then evaluated using techniques developed by machine learning practioners in the past. For example, performance is often measured by taking a model and a defined prompt, and then using this combination to process a test set of known documents and then measuring how close to a target or expectation the output from the prompted LLM is (Figure 1).

This approach is prone to the same sort of issues that were observed in previous machine learning systems. Arbitrary decisions can embed bias, create blindspots, and generally distort performance. Optimisations with frameworks like DsPy [1] can create overfitting, where a selected prompt and model combination is optimised to the point that it becomes brittle and fails to work well on examples outside of the test set. Another issue is with the type of output that an LLM typically gen-

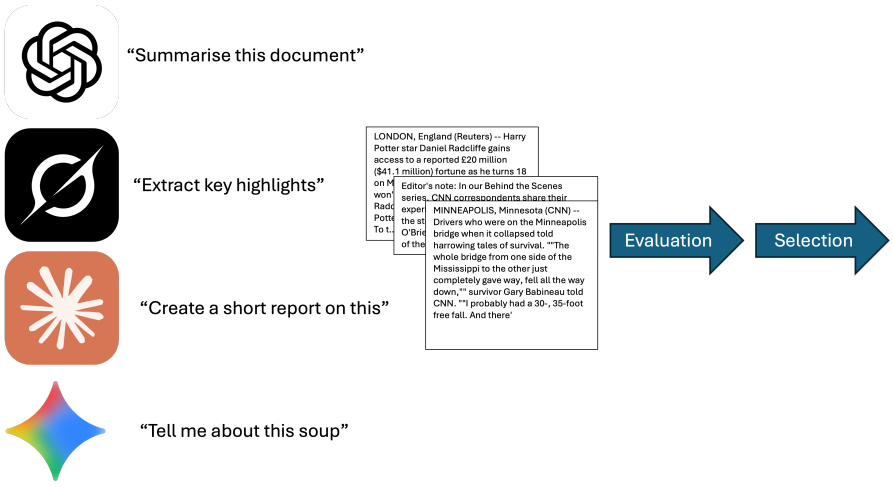

Figure 1: Modern AI : combining generic LLMs with specific prompts to process target data. Evaluation is via a test set using some fitness measure.

erates. Typically machine learning models would produce classification (cat | dog | cow | sheep) or decision (yes | no) outputs, but an LLM will typically be prompted to produce long strings of text. Additionally, LLMs are engineered in such a way that they can produce inconsistent results when passed the same inputs.

So, the feedback generated by running examples over a test set is much less useful for an LLM application than it is for evaluating a traditional machine learning model, although it was problematic for that as well.

2 What does useful look like?

The questions that stakeholders ask about AI applications are both simple and reasonable.

- "Will this work in production?"
- "Is this as good as it could be?"
- "How much should we spend on improving this?"

Our intuition was that rather than evaluating a wider range of test cases it would be more effective to evaluate existing test cases on a wider range of prompts.

3 Isophore: the LLM testing engine

Isophore is an engine that takes a prompt and an LLM and tests it to… well not quite destruction; instead it tests it into a confession. The confession that Isophore creates contains all the information from the local manifold (fitness landscape) around your prompt. Then Isophore squeezes this information for statistical insights about how the LLM is reacting to your prompt and data. These stats are

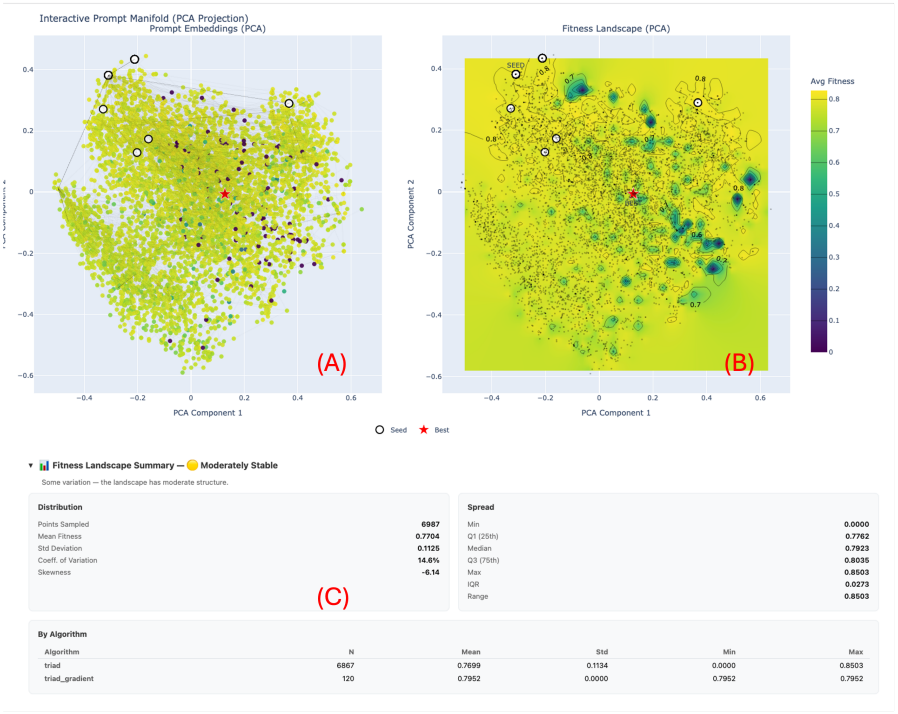

Figure 2: Typical output from a run of Isophore showing the manifold visualisation (B) and interactive investigation window (A) and the statistical summary of the manifold (C).

then turned into the insight that can help you to qualify your AI system for investment, production, or disposal.

An early view of what Isophore provides is shown in figure 2. The labels (A), (B), (C) are not part of the system output but are shown to allow a better description of the functionality. Window (A) shows all of the points that Isophore has sampled in its attempt to map the manifold for the target prompt. This is an interactive display that allows the user to explore the outcome of the searching prompt by prompt and shows the fittest prompt that was found during the search. (B) shows the fitness landscape on the manifold, essentially the map of the space around the users prompt. (C) shows the statistical view of the landscape, including qualitative feedback about what's been discovered, in this case a moderate, benign structure was discovered around the prompt which provides reassurance that while there might be more optimal candidates for the prompt in the local landscape they are unlikely to represent genuinely better and more robust approaches than the one that the user selected initially.

4 Isophore Advantages

What are the technical developments that enable this kind of mapping.

- Proprietary search algorithm; common art is to use HMCTS [2] to explore

the manifold, but this does not account for the variability of LLM response, and does not have the appropriate properties to find areas of high variability – rather it aims to explore areas of failure. Our alternative searches for and then explores intensively areas of significant fitness variability and shares information from multiple search threads to achieve this.

- Statistical & manifold interpretation; rather than discovering failure Isophore develops an informative map that guides a business decision (develop, deploy, dispose).
- Visual interpretation; by providing intuitive visualisations Isophore enables users to rapidly make judgments about the qualities of the landscapes that they are using. Significant statistical innovation was required to manage the plotting of PCA based visualisations that remain consistent between runs

5 Next Steps.

Isophore is under development. Currently we have explored a small fraction of the potential machinery that could be brought to bear on this problem. Our next steps will explore alternative fitness measurement mechanisms, better manifold search and stronger statistical interpretation of the landscape.

References & Further Reading

[1] Omar Khattab et al. *DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines*. 2023. URL: https://github.com/stanfordnlp/dspy.

[2] Yue Huang et al. *ProbeLLM: Automating Principled Diagnosis of LLM Failures*. 2026. arXiv: 2602.12966 [cs.CL]. URL: https://arxiv.org/abs/2602.12966.