Contribution ID: **44** Type: **not specified**

# Selecting Statistical Metrics for Prompt Optimization of Locally Hosted LLMs: A Case Study in Asset Registration

In a client project focused on asset registration, we transitioned from a publicly available LLM (Gemini) to a locally hosted LLM to prevent the potential leakage of sensitive manufacturing and customer data. Due to hardware constraints (GPUs with a maximum of 12 GB VRAM), the performance of local LLMs was initially inferior to hosted models. Therefore, prompt optimization became crucial to achieve comparable output quality. We explored the application of LLM-specific statistical evaluation metrics —G-Eval, Answer Relevance, Prompt Alignment, Sensitivity, and Consistency —using the Langfuse framework deployed on-premises. We evaluated three quantized models (llama3.1 8b, mistral 7b , deepseek-llm 7b) on a dataset of 200 simulated asset registration requests in multiple languages (German, English, Turkish), reflecting real-world operational complexity.
Reference values for evaluation were manually extracted by human coders. Prompt engineering techniques, including Zero-shot, One-shot, and Few-shot strategies, were applied systematically. Only information present in the registration prompts was extracted; no hallucinations occurred (temperature was fixed at 0). However, significant effort was needed to maximize information extraction through tailored prompts.
This case study demonstrates how statistical metrics can guide and validate prompt optimization, especially under constrained computing conditions and multilingual input. We invite discussion on alternative metrics, further application domains, and lessons learned from fitting models primarily trained for English to German and mixed-language tasks.

## Special/ Invited session

## Classification

Mainly application

## Keywords

Statistical Evaluation Metrics; Prompt Optimization; Locally Hosted LLMs

**Primary author:** Dr MOESER, Guido (masem research institute)

**Co-author:** Mr DUB, Igor (Referat Smart City Wiesbaden)

**Presenter:** Dr MOESER, Guido (masem research institute)

**Track Classification:** AI: Interpretability and Trustworthiness