## Bernardo Aceves Partida

Automating Data Management Plan Evaluation Through the Integration of Machine-Actionable Data Management Plans and Fair Implentation Profiles

**Kriterium 1 – Problem Statement**  
The problem is clearly described by identifying inefficiencies in current research data management practices and the lack of an automated evaluation framework. However, the statement would benefit from explicit, measurable KPIs to quantify the current issues.

**Kriterium 2 – Goals and Expected Outcome**

The goals and expected outcomes are well articulated, detailing a conceptual framework, mapping model, and evaluation system. Enhancing the proposal with concrete numerical targets or baseline metrics would strengthen its impact.

**Kriterium 3 – Research Questions**  
The research questions are precise and aligned with the thesis objectives. Clarifying what constitutes “optimal” FAIR assessment in measurable terms could further refine the scope.

**Kriterium 4 – Research Methods**  
The step-by-step methodology—from literature review through to testing—is appropriate and coherent. Explicitly linking each methodological step to established frameworks or seminal papers would enhance the scientific rigor.

**Kriterium 5 – Evaluation**  
The proposed evaluation approach, involving comparisons with expert assessments and use of FAIR metrics, is promising. Detailing specific evaluation metrics and benchmarks would improve clarity regarding how success is quantitatively determined.

**Kriterium 6 – State of the Art**  
The literature review demonstrates a solid grasp of current approaches and identifies a clear gap. A brief discussion of limitations or potential biases in the existing literature could provide additional depth.

**Kriterium 7 – Relevance to the Curriculum**

The proposal effectively connects its subject matter to multiple relevant courses, thereby underlining its academic and practical relevance within the curriculum.

**Kriterium 8 – Overall Assessment**  
The document is comprehensive and well-structured with a clear argumentative line. While the content is rich, some sections could be more concise, and the inclusion of precise, quantifiable metrics would enhance its overall clarity.

**Kriterium 9 – Miscellaneous**  
The proposal is innovative and addresses a timely research gap. Minor revisions, particularly in detailing quantitative aspects and linking methods to established research, would further solidify its foundation.

**Gesamtfeedback**  
Overall, this thesis proposal presents a well-founded and innovative approach to automating DMP evaluation by integrating machine-actionable plans with FAIR Implementation Profiles. The methodology is robust and the research questions are well chosen, demonstrating a clear understanding of current challenges in research data management. With minor refinements—especially in incorporating concrete metrics and explicitly referencing established methods—the proposal has strong potential to make a significant contribution to the field.

**Questions**

1. How will you quantitatively benchmark the current inefficiencies in research data management to establish a robust baseline for your evaluation?
2. Which specific KPIs do you plan to employ to measure the FAIRness of DMPs, and how will these metrics be validated against expert assessments?
3. What potential challenges do you anticipate when aligning FIPs with maDMPs, and what strategies do you propose to mitigate them?
4. In what ways does your methodological approach distinguish itself from existing automated evaluation tools, especially for early-stage DMPs?

## Stipe Babic

Automated Root-Cause Analysis for Unexpected System Behavior

**Kriterium 1 – Problem Statement**  
The problem is clearly outlined by emphasizing the need for an automated root-cause analysis tool to reduce downtime and repair costs in agricultural tractors. However, while the impact (financial, operational, safety) is discussed, the statement would benefit from specifying measurable KPIs to gauge current inefficiencies.

**Kriterium 2 – Aim of the Thesis and Expected Results**  
The aim is well defined: to develop a tool that identifies and explains root causes of failures in tractors using both baseline statistical methods and more advanced techniques. Expected results are practical and applicable. Explicit success metrics—such as error detection accuracy or reduction in repair time—would strengthen this section.

**Kriterium 3 – Research Questions**  
Although explicit research questions are not separately listed, the proposal implies key inquiries around error identification, causality, and the integration of multiple data sources. Stating clear, measurable research questions would enhance the focus and evaluation of the investigation.

**Kriterium 4 – Research Methods**  
The methodology is systematically organized into data exploration, literature review, and implementation phases. The approach is appropriate, yet the proposal could be improved by explicitly linking each methodological step to established literature (e.g., association rule mining, causal inference, LLM applications).

**Kriterium 5 – Evaluation**  
The evaluation strategy hints at comparing baseline statistical methods with more sophisticated techniques and cross-referencing with domain knowledge. However, a more detailed description of evaluation metrics and benchmarks—such as accuracy, precision, or recall—is needed to clarify how success will be measured.

**Kriterium 6 – State of the Art**  
The state of the art is comprehensively reviewed, with references to relevant studies and methods in RCA, causal inference, and LLM-based approaches. The discussion effectively positions the proposed work within current research, though a brief mention of potential limitations in existing methods would provide additional context.

**Kriterium 7 – Relevance to the Curriculum**  
The proposal clearly demonstrates relevance to the Data Science curriculum by linking it to key courses such as Natural Language Processing, Statistical Computing, and Generative AI. This connection underlines the academic and practical significance of the research.

**Kriterium 8 – Overall Assessment**  
The proposal is well-structured and thorough, balancing practical application with theoretical exploration. It could be further enhanced by incorporating explicit research questions, more concise sections, and clearly defined quantitative benchmarks for evaluation. The writing is clear and professional, supporting a coherent line of argument.

**Kriterium 9 – Miscellaneous**  
The proposal is innovative, particularly in its integration of heterogeneous data sources and advanced techniques such as LLMs. A deeper discussion on the challenges of merging diverse data types (time series, workshop logs, domain knowledge) and the limitations of the chosen methods would add further depth.

**Gesamtfeedback**

Overall, this thesis proposal presents a promising approach to automating root-cause analysis for agricultural tractors. The problem is both timely and significant, and the proposed methodology is sound, integrating conventional statistical analysis with advanced AI methods. With minor enhancements—such as explicitly stated research questions, quantifiable evaluation metrics, and a discussion on potential data integration challenges—the proposal is well positioned to make a substantial contribution to the field of automated system diagnostics.

**Questions for the Author:**

1. Can you elaborate on the specific KPIs you plan to use for evaluating the RCA tool, and how these will quantitatively demonstrate improvements over current practices?
2. How do you intend to integrate the heterogeneous data sources (time series, workshop logs, and domain knowledge) to ensure a cohesive and effective analysis?
3. What potential limitations do you foresee in applying LLMs to this domain, and what strategies might you employ to mitigate these challenges?
4. Would you consider explicitly formulating research questions or hypotheses to guide your methodology, and if so, how will they shape your experimental design?

## Tilman Kerl

Evaluation of Sparse Autoencoder-based Refusal Features in LLMs: Dataset-dependence study

**Kriterium 1 – Problem Statement**  
The proposal clearly defines the challenge of ensuring safe refusal behavior in LLMs by leveraging SAE-based feature extraction, emphasizing the influence of training data on feature quality. It would be beneficial to specify measurable KPIs for assessing refusal performance and robustness.

**Kriterium 2 – Goals and Expected Outcome**  
The objectives are well articulated: to evaluate the impact of dataset characteristics on SAE-extracted refusal features and to assess their effect on model safety and downstream performance. Including explicit numerical targets or success thresholds would further strengthen this section.

**Kriterium 3 – Research Questions**  
The three research questions are clearly formulated and directly address the core issues—dataset influence, predictive dataset properties, and performance comparisons across different training corpora. They provide a solid basis for empirical investigation.

**Kriterium 4 – Research Methods**  
The methodological approach is comprehensive, covering model selection, SAE training, dataset collection, feature extraction, and steering. The step-by-step plan is sound, though clarifications on baseline comparisons and integration of heterogeneous evaluation methods could enhance clarity.

**Kriterium 5 – Evaluation**  
The evaluation plan is robust, divided into phases assessing SAE quality, refusal behavior (refusal and over-refusal rates), and downstream performance (using benchmarks like MMLU and JailbreakBench). More detailed descriptions of evaluation metrics and thresholds would help in quantifying improvements.

**Kriterium 6 – State of the Art**  
The literature review is extensive and well-grounded, covering both traditional fine-tuning approaches and innovative SAE-based methods. It successfully positions the proposed work within current research while highlighting gaps regarding dataset dependence and feature disentanglement.

**Kriterium 7 – Relevance to the Curriculum**  
The proposal demonstrates clear relevance by linking its topics to core courses in Machine Learning, NLP, Deep Learning, and Information Retrieval. This connection supports both the theoretical and practical aspects of the study within the curriculum framework.

**Kriterium 8 – Overall Assessment**  
Overall, the proposal is thorough and technically sophisticated. It successfully integrates theoretical foundations with a practical implementation plan. Some sections could be made more concise, and clearer quantifiable metrics would enhance its overall impact.

**Kriterium 9 – Miscellaneous**  
The proposal is innovative and timely, addressing critical issues in LLM safety. Minor clarifications regarding the integration of feature steering and baseline comparisons, as well as the management of variability across datasets, would further refine the study.

**Gesamtfeedback / Rückmeldung an den / die Autor/in**  
In summary, this thesis proposal presents a strong and well-researched approach to improving LLM safety by evaluating SAE-based refusal features. The research questions are clear and the methodological framework is comprehensive, effectively bridging theoretical insights with practical evaluation. Clarifying quantitative performance indicators and integrating some of the methodological details more explicitly would enhance the proposal's clarity and impact. Overall, the proposal shows great potential for advancing mechanistic interpretability and safety in language models.

**Questions for the Author:**

1. Can you specify which quantitative metrics or KPIs you will use to measure the robustness and interpretability of the SAE-extracted refusal features?
2. How will you handle variability in dataset characteristics when comparing the impact of different training corpora on feature extraction?
3. Could you clarify how you plan to integrate and compare results from SAE-based methods with baseline models, particularly regarding downstream performance and safety alignment?
4. What strategies will you employ to ensure that manipulating refusal-related features does not negatively impact the model’s performance on tasks unrelated to refusal behavior?