Topic	**Systematic Identification and Filtering of Conflicting Datasets**
Overview	Large Language Models (LLMs) like GPT and BERT have transformed how we interact with AI by understanding and generating human-like text. Initially trained on extensive datasets to grasp a wide range of language nuances, LLMs often require fine-tuning to adapt to specific tasks or behaviors in targeted domains. This fine-tuning process involves training the model on a smaller, specialized dataset that closely represents the scenarios it will face, enhancing its ability to perform designated tasks accurately and interact in expected ways. Fine-tuning aligns the model's responses with user intents and desired outcomes, ensuring the model's interactions are both relevant and engaging. This process is critical in making LLMs effective and reliable in practical, real-world settings. The performance of a fully trained model is highly dependent on the quality of the used data. This thesis aims to devise and refine methods for the systematic identification and resolution of conflicts within datasets. That is especially important for a fine-tuned language model, to distinguish between using "trained knowledge" and retrieval mechanisms. In this thesis you will address how conflicting data entries can compromise the accuracy of machine learning models and retrieval systems. The thesis will focus on developing and assessing effective strategies for deduplication and filtering of these datasets to enhance their reliability and performance. **Objectives:** • Literature Review: Review existing research on data conflicts and deduplication in datasets used for machine learning and information retrieval. Examine techniques and frameworks currently employed to handle conflicting data. • Method Development: Develop innovative methods and algorithms for detecting and resolving data conflicts. This includes designing or improving upon existing deduplication techniques. • Dataset Compilation: Compile a dataset that includes potential conflicts specifically tailored to test the developed methods. This dataset should be representative of real-world scenarios

	where knowledge-based systems and retrieval systems are applied. • Evaluation: Assess the effectiveness of the methods by examining their behavior, particularly in scenarios where they are required to utilize mechanisms for retrieving additional external information.
Language	German or English
Exemplary literature	Kortukov, E., Rubinstein, A., Nguyen, E., & Oh, S. J. (2024). Studying Large Language Model Behaviors Under Realistic Knowledge Conflicts (Version 1). arXiv. Zhuang, Y., Yu, Y., Wang, K., Sun, H., & Zhang, C. (2023). ToolQA: A Dataset for LLM Question Answering with External Tools. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), Advances in Neural Information Processing Systems (Vol. 36, pp. 50117–50143).
Additional information	Type of thesis: Bachelor or master thesis Start: As soon as possible Requirements: - Programming skills, preferably in Python Publication: Excellent work can be submitted and published as a conference paper or at TU Prints
Contact	If you are interested, please contact Nils Schönfeld (nilslucas.schoenfeld@bwl.tu-darmstadt.de) Prof. Dr. Dr. Ruth Stock-Homburg