Hochschule Neu-Ulm
University of Applied Sciences

IVECO ACADEMY

Master Thesis

in fulfillment of the requirements for the degree of

Master of Science in

Artificial Intelligence and Data Analytics

at **Hochschule Neu-Ulm**

From Tickets to Predictive Training: An AI Framework for Organizational Training-Gap Detection

---

submitted by

Mohamed Mahfoudh Bouna

Matriculation No.: 376295

Submitted on 04.03.2026

Academic Supervisors

Prof. Dr. Jörg-Oliver Vogt

Prof. Dr. Marten Risius

Information Management

Hochschule Neu-Ulm

Company Supervisor

Severin Tauber

Head of IVECO Academy

IVECO Deutschland AG

Declaration of Authorship

I hereby declare that the Master's thesis titled *"From Service Tickets to Predictive Training: An AI Framework for Organizational Training-Gap Detection"* is my own work. All sources and aids used have been indicated as such. All passages quoted from publications or paraphrased from publications have been attributed to their sources.

This thesis has not been submitted to any other institution for the purpose of obtaining an academic degree or qualification.

Ulm, 01 March 2026

Mohamed Mahfoudh Bouna

Abstract

IVECO Academy Germany currently lacks a systematic, data-driven method to detect training gaps from operational support data. This thesis applies Design Science Research to design, implement, and evaluate KI-Academy, a hybrid AI framework for organizational skill-gap detection. The system processes 10,305 German-language Technical Hotline Desk (THD) tickets alongside 185 training courses across three analytical pipelines: Pipeline A (SBERT embeddings and HDBSCAN clustering), Pipeline B (a locally deployed large language model via Ollama using qwen2.5:7b), and Pipeline C (a hybrid integration of A and B with Retrieval-Augmented Generation). The framework produced 14 topic clusters with 90% ticket coverage, 15 gap classifications, including 11 kein_gap, 2 komplett_neu, and 2 inhaltlich_teils_offen, 14 course recommendations, and 4 proposals for entirely new training content. Expert validation by five IVECO Academy staff members confirmed the system's practical relevance and its suitability as a foundation for future AI-based training planning.

Keywords: Training-Gap Detection · NLP · LLM · Hybrid AI Pipeline · RAG

Zusammenfassung

Der IVECO Academy Deutschland fehlt bislang eine systematische, datengetriebene Methode zur Erkennung von Trainingslücken aus operativen Supportdaten. Diese Masterarbeit wendet Design Science Research an, um das hybride KI-Framework KI-Academy zu entwerfen, zu implementieren und zu evaluieren. Das System verarbeitet 10.305 deutschsprachige Tickets des Technical Hotline Desk (THD) sowie 185 Trainingskurse in drei Analysepipelines: Pipeline A (SBERT-Embeddings und HDBSCAN-Clustering), Pipeline B (ein lokal betriebenes LLM via Ollama mit qwen2.5:7b) und Pipeline C (hybride Integration von A und B mit RAG). Das Framework erzeugte 14 Themencluster mit 90% Ticketabdeckung, 15 Lückenklassifikationen, darunter 11 kein_gap, 2 komplett_neu und 2 inhaltlich_teils_offen, , 14 Kursempfehlungen sowie 4 Vorschläge für vollständig neue Trainingsinhalte. Die Expertenvalidierung durch fünf Mitarbeitende der IVECO Academy bestätigte die praktische Relevanz und die Eignung des Systems als Grundlage für zukünftige KI-gestützte Trainingsplanung.

Schlüsselwörter: Trainingslückenanalyse · Natürliche Sprachverarbeitung · Große Sprachmodelle · Hybride KI-Pipeline · RAG

List of Abbreviations

Abbreviation	Full Term
ADAS	Advanced Driver Assistance Systems
API	Application Programming Interface
BPMN	Business Process Model and Notation
CSV	Comma-Separated Values
DSR	Design Science Research
DSRM	Design Science Research Methodology
EATS	Exhaust Aftertreatment System
FEDS	Framework for Evaluation in Design Science Research
GDPR	General Data Protection Regulation
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
HRD	Human Resource Development
KI	Künstliche Intelligenz (Artificial Intelligence)
LDA	Latent Dirichlet Allocation
LLM	Large Language Model
NLP	Natural Language Processing
OTA	Over-the-Air (software update)
RAG	Retrieval-Augmented Generation
RLHF	Reinforcement Learning from Human Feedback
SBERT	Sentence-BERT
TF-IDF	Term Frequency–Inverse Document Frequency
THD	Technical Hotline Desk
TNA	Training Needs Analysis
WEF	World Economic Forum

List of Figures

Contents

1 Introduction

1.1 Background and Motivation

Every day, the Technical Hotline Desk (THD) of IVECO, a major European commercial vehicle manufacturer, receives tens or hundreds of support tickets from service technicians across its dealer network, accumulating to more than 10,000 tickets per year for the German market alone. Each ticket documents a vehicle fault, a diagnostic challenge, or a technical problem that the technician could not resolve without external assistance. Taken in aggregate, this continuous stream of structured and unstructured operational data constitutes one of the richest available records of where employees lack the knowledge or skills to perform independently. Despite this potential, THD ticket data has historically served a single operational purpose: issue resolution and workload management. It has never been systematically connected to workforce training planning.

The research presented in this thesis originated from precisely this observation. Management at IVECO Academy Germany, the organizational unit responsible for designing and delivering technical training programs for service technicians across the IVECO vehicle portfolio, identified the absence of a data-driven method for detecting training gaps as a pressing organizational challenge. This practitioner-initiated framing ensures that the research problem is real and consequential, not constructed for academic convenience. The author's dual role as an IVECO employee and master's student provided both the institutional access and the organizational grounding necessary to pursue this problem as a rigorous Design Science Research study.

The current approach to training needs analysis (TNA) at IVECO Academy Germany relies on expert interviews, manager observations, and periodic technician surveys, all of which are inherently reactive, subjective, and slow. By the time a skills deficit surfaces through these channels, it has frequently already manifested as repeated escalations, increased support costs, or vehicle downtime. This lag is not merely an organizational inconvenience; it represents a structural misalignment between the pace at which vehicle technology evolves and the pace at which training curricula can respond. The introduction of new powertrains, advanced telematics systems, and complex aftertreatment technologies in modern commercial vehicles has accelerated this misalignment considerably. Training programs that were current eighteen months

ago may no longer address the diagnostic competencies that technicians most urgently need today.

The broader macroeconomic context reinforces the urgency of this problem. (World Economic Forum, 2023) projects that six in ten workers will require significant upskilling by 2027, with technology-intensive industries facing the steepest competency gaps. In sectors such as commercial vehicle manufacturing, where product complexity grows with each new vehicle generation, organizations that can identify emerging skill gaps proactively, before they escalate into operational failures, gain a measurable competitive and operational advantage. Reactive TNA processes cannot meet this standard.

The technical preconditions for a data-driven alternative now exist. Advances in natural language processing (NLP), transformer-based language models, and topic modeling techniques have created a rich toolkit for extracting structured intelligence from unstructured text at scale. Concurrently, the emergence of locally deployable large language models (LLMs) has made it feasible to apply powerful generative reasoning capabilities within enterprise environments where data governance and privacy requirements prohibit the use of cloud-based API services. These developments open a genuine possibility: using the very ticket data that documents technician struggles as the primary input for automated, curriculum-aligned training gap detection.

Yet, despite these technical advances and the clear organizational need, no prior work has combined classical NLP pipelines and locally deployed LLMs into a hybrid framework specifically designed to map ticket topics to training curriculum gaps in a privacy-preserving, enterprise-ready architecture. The data exists. The methods exist. The organizational challenge exists. What has been missing is the integrated system and the design knowledge required to bring them together. This thesis addresses that gap.

1.2 Problem Statement and Research Objectives

Three interrelated problems motivate this research. The first is the structural disconnect between service operations and workforce development. THD tickets are a direct, continuous, and granular record of technician knowledge deficits, yet as (Ferreira & Abbad, 2013) demonstrated in their systematic review of 122 TNA studies, training needs assessment continues to rely predominantly on subjective instruments such as interviews, questionnaires, and managerial judgment. These approaches are inherently

lagging: they identify gaps after they have become visible enough to surface through human observation. The result is a reactive training cycle that cannot keep pace with the rate of technical change in modern vehicle systems. IVECO Academy Germany's current TNA process reflects this pattern precisely, making the organizational context a compelling and generalizable instance of a well-documented structural problem.

The second problem is the absence of a scalable, privacy-preserving AI method for this specific purpose. While NLP and LLM techniques have advanced substantially, their application to organizational TNA remains nascent. Existing work in service analytics has applied topic modeling to IT helpdesk and software bug report data, but there is no prior work that combines classical NLP clustering with LLM-based reasoning in a hybrid architecture designed explicitly for technical support ticket data in an automotive enterprise context. The locally deployable constraint, necessitated by IVECO's data governance requirements, which preclude the transmission of operational data to external cloud services, further narrows the design space in a direction that existing literature does not address.

The third problem is a methodological design question: it remains unclear which analytical approach offers the best trade-off between analytical quality, computational efficiency, and enterprise integration suitability. A purely classical NLP pipeline offers reproducibility and interpretability but may lack the semantic reasoning capacity required to assess curriculum relevance. A purely LLM-based approach offers flexible reasoning but introduces risks related to computational cost and output consistency. A hybrid architecture combining both raises questions about how the components should be integrated and how their respective contributions should be evaluated. Resolving this design question requires not only implementing all three configurations but comparing them systematically under identical conditions.

Against this background, this thesis pursues four research objectives. The primary objective is to design and implement a functional AI framework that transforms raw, anonymized THD ticket data into structured, curriculum-aligned training gap intelligence. The second objective is to implement and compare three pipeline configurations (classical NLP-only, LLM-only, and hybrid) within the same integrated system and under identical input conditions, enabling a controlled comparison of their analytical output. The third objective is to evaluate the framework through structured expert validation with IVECO Academy Germany staff, grounding the assessment in practitioner judgment. The fourth objective is to produce transferable design knowledge

that extends beyond the IVECO context and is applicable to other organizations facing analogous challenges in skill-gap detection from operational data.

These objectives yield three distinct contributions. Practically, the thesis delivers a fully implemented, locally deployable AI framework, validated with domain experts. Methodologically, it provides a Design Science Research-based evaluation design comparing NLP and LLM pipeline components within a single integrated system. Academically, it produces design knowledge on hybrid NLP–LLM architectures for organizational training gap detection, a contribution that qualifies, in the sense of (Gregor & Hevner, 2013), as an 'Invention': a novel artifact addressing a problem for which no prior solution exists in the literature.

1.3 Research Question and Scope

The research question guiding this study is: "What AI-based framework can be designed and evaluated to identify training gaps from technical support ticket data, and how do classical NLP pipelines and locally deployed large language models compare in effectiveness and integration suitability for this purpose?"

This question embeds two complementary imperatives. The first is constructive: it calls for the design and implementation of an artifact. The second is evaluative: it demands a comparative assessment of the artifact's constituent approaches. Together, they define a research agenda that is fully aligned with the Design Science Research (DSR) paradigm as articulated by (Hevner et al., 2004), which holds that IS research should produce utility-bearing artifacts alongside the propositional knowledge necessary to understand their design and performance.

The scope of this study is defined by a set of explicit boundaries that both enable rigorous execution and must be acknowledged as constraints on generalizability. The study is conducted within IVECO Academy Germany, a single organizational unit of a large European commercial vehicle manufacturer, and the ticket corpus consists exclusively of anonymized, German-language THD tickets from this context. Findings are therefore subject to confirmation in other organizational settings before broad generalization. The LLM component is deployed locally on-premise via Ollama; cloud-based LLM APIs such as those provided by OpenAI or Google are explicitly out of scope, in compliance with IVECO's data governance requirements. The retrieval-augmented generation (RAG) mechanism used to ground LLM outputs in curriculum documents

follows the architecture introduced by (Lewis et al., 2020), adapted for local deployment.

The research follows a DSR methodology; its goal is to produce a functional artifact and associated design knowledge, not to develop a predictive statistical model or conduct a controlled experiment in the social-scientific sense. Evaluation relies on structured expert judgment from internal IVECO Academy Germany staff; large-scale user studies are out of scope. The framework is designed to produce training gap intelligence as decision support for human training planners. Autonomous training planning decisions, in which the system acts without human review, are explicitly out of scope and would require a separate line of research addressing questions of organizational authority, accountability, and model reliability that lie beyond the boundaries of this study.

1.4 Thesis Structure

This thesis moves from theoretical foundation through system design and empirical validation to synthesized design knowledge and conclusion. Chapter 2 reviews the relevant literature across six domains (organizational learning and TNA, NLP in service analytics, classical NLP techniques, LLMs in enterprise contexts, hybrid architectures, and Design Science Research), establishing the conceptual and technical foundations upon which the framework is built. Chapter 3 details the research design and DSR methodology, including the evaluation strategy and the protocol used for expert validation. Chapter 4 presents the system architecture and the implementation of all three pipeline configurations. Chapter 5 reports the empirical results of the comparative pipeline analysis. Chapter 6 evaluates the framework against the research objectives and expert criteria. Chapter 7 synthesizes the design knowledge produced, reflects on limitations, and identifies directions for future work before concluding the thesis.

2 Literature Review and Theoretical Foundation

2.1 Enterprise Knowledge Management and Training Needs Analysis

In knowledge-intensive industries where the competence profile of field personnel is continuously reshaped by new vehicle platforms, emission legislation, and electronic architectures, the systematic identification of workforce skill gaps constitutes a fundamental operational necessity. (Noe, 2020) positions training needs analysis (TNA) as the non-negotiable prerequisite for evidence-based training strategy, and (Goldstein & Ford, 2002) operationalized TNA through the canonical three-phase model, organizational analysis, task analysis, and person analysis, that provides the conceptual architecture the thesis's AI framework is designed to automate.

Despite its theoretical standing, TNA in organizational practice remains persistently reactive and subjective. (Ferreira & Abbad, 2013), reviewing 122 empirical TNA studies, found a near-total absence of automated or data-driven alternatives to episodic expert interviews and perception-based surveys, a condition directly observable in the IVECO dealer network prior to this thesis's intervention, where gap identification was driven entirely by managerial perception rather than operational evidence. (Gubbins et al., 2018) reinforced this critique from an evidence-based HRM perspective, demonstrating that practitioner reliance on subjective judgment introduces systematic bias that distorts gap assessments and must be replaced by systematically gathered data. The organizational consequence is that skill gaps typically surface only after they have already produced service failures, a vehicle delayed unnecessarily in the workshop, a fault misdiagnosed, customer satisfaction eroded, by which point the training response is remedial rather than preventive. (Kraiger, 2014) and (Marsick & Watkins, 2003) anticipated the remedy: learning analytics and embedded knowledge-capture mechanisms can transform TNA from an episodic activity into a continuous, data-grounded process, a transformation the thesis operationalizes by embedding an AI pipeline into IVECO's existing ticket management system.

The theoretical basis for treating support tickets as a valid TNA input is provided by the informal learning literature. (Manuti et al., 2015) demonstrated that the majority of organizationally relevant learning is informal and generates textual traces in operational information systems, while (Cerasoli et al., 2018), meta-analyzing 79 studies, confirmed that such informal signals carry strong diagnostic validity as indicators of unmet competence needs. A technician's support ticket, encoding the vehicle model, reported symptom, and fault system that exceeded diagnostic capability, is precisely such a signal: it documents an encounter with knowledge the technician demonstrably lacked at the time of service. (Capaldo et al., 2006) further argued that accurate competency diagnosis requires contextual, situational data rather than generic role descriptions, and an aggregated ticket corpus provides exactly the continuously updated, context-rich competency map that situationalist theory demands. (Ismail, 2016) and (Pham et al., 2019) supplied cross-domain empirical evidence that training programs calibrated to actually identified gaps rather than assumed needs produce measurably superior organizational outcomes, while (Nguyen et al., 2023) confirmed that technology-based knowledge-sharing artifacts, of which the thesis's gap-detection dashboard is an instance, produce measurable improvements in employee performance. Taken together, this body of evidence establishes that technical support tickets are a legitimately rich, underexploited signal of operational skill gaps, and that automating their analysis for TNA purposes is both theoretically grounded and organizationally necessary.

2.2 Text Mining and Topic Modeling in Service Analytics

2.2.1 Prior Work on Support Ticket Clustering

The existing NLP literature on technical support ticket data is substantial but has without exception pursued exclusively operational objectives, routing, triage, severity classification, and resolution-time prediction, without connecting discovered textual patterns to workforce training or curriculum design. (Zangari et al., 2023), the most comprehensive current survey, benchmarked transformer-based classifiers on real-world IT ticket datasets and found F1 scores of only 38–55%, establishing both the practical difficulty of automated ticket interpretation and the critical importance of domain-adapted

preprocessing and embedding strategies. The operational orientation is consistent across prior work: (Tian et al., 2015) demonstrated that full ticket text analysis outperforms metadata-based classification, methodologically justifying the thesis's decision to analyze complete ticket text rather than pre-existing categorical fields; (Muriana & Vizzini, 2017) showed that text mining of engineering incident records surfaces recurring operational risk patterns and informs proactive organizational responses, providing industrial precedent for extracting actionable signals from accumulated technical records; and (Chen et al., 2019) addressed intent classification and response generation in technical support dialogue, again, entirely operational in purpose and scope. The disciplinary boundary between operational service analytics and organizational learning is, in the published literature, entirely untraversed.

The closest methodological analogue to the thesis is (Rosen & Shihab, 2016), who applied LDA topic modeling to 50,000 Stack Overflow developer questions to identify recurring knowledge gaps. Their approach, aggregating unstructured technical queries, discovering thematic clusters, and interpreting them as evidence of systematic deficits, mirrors the thesis's pipeline conceptually. Critically, however, their conclusions were directed toward API design rather than training planning, and the conceptual step from ticket cluster to actionable training recommendation is absent from their study as from every other work in the ticket analytics literature. (Zanker et al., 2019) and (Lin et al., 2002) offer complementary framing, positioning the inference of latent needs from observable text signals as analogous to recommendation systems, a framing that situates the thesis as a training recommendation engine driven by operational text.

(Efstathiou et al., 2018) demonstrated that generic embedding models perform significantly worse than domain-adapted alternatives on specialized technical vocabulary, directly motivating the thesis's use of a multilingual model. (Conneau et al., 2020), whose XLM-RoBERTa established high-quality multilingual representation learning at scale, provides the architectural lineage of the multilingual SBERT variant deployed here, a design necessity for a corpus written predominantly in German with dense automotive terminology.

2.2.2 Topic Modeling Techniques and Their Limitations

LDA (Blei et al., 2003) remains the canonical baseline for unsupervised topic discovery and the primary methodological reference against which the thesis's clustering approach

must be positioned. Its assumption that each document is a mixture of all topics makes it theoretically flexible but poorly matched to short, focused support tickets that typically address a single technical problem. (Röder et al., 2015) established C_V coherence as the most reliable automated proxy for human-perceived topic quality, the evaluation standard adopted by this thesis, building on the mutual-information foundation of (Newman et al., 2010); both frameworks are applied directly in the thesis's quantitative NLP assessment. (Agrawal et al., 2018) demonstrated that adaptive, data-driven hyperparameter selection consistently outperforms fixed defaults on technical corpora, justifying the thesis's strategy of computing HDBSCAN parameters dynamically from embedding space geometry rather than specifying them manually. The fundamental limitation of all bag-of-words models for this application is their inability to capture semantic equivalence across different surface forms: two tickets describing a diesel particulate filter regeneration failure may share no vocabulary whatsoever yet are conceptually identical, a lexical gap that no term co-occurrence statistic can bridge. Dense sentence embeddings resolve this by representing semantic content directly in continuous vector geometry, enabling similarity computations that reflect genuine conceptual proximity regardless of surface variation, which is why the embedding-based approach adopted in this thesis supplants LDA-style methods for the ticket clustering task.

2.3 Classical NLP Techniques for Document Analysis

2.3.1 Text Preprocessing, Embeddings, and Clustering

The representational evolution from TF-IDF to dense contextual sentence embeddings is the foundational technical advance enabling semantic clustering of short technical texts. (Mikolov et al., 2013), (Vaswani et al., 2017), and (Devlin et al., 2019) established, respectively, the distributional embedding paradigm, the Transformer self-attention architecture, and BERT as the bidirectional pre-trained foundation from which all downstream sentence embedding architectures derive. (Reimers & Gurevych, 2019) addressed BERT's weakness at sentence-level representation by introducing SBERT, which produces semantically rich, efficiently comparable sentence embeddings through

Siamese network training on natural language inference data, enabling practical large-scale clustering of ticket corpora that would be computationally prohibitive with naive BERT token averaging. (Lau & Baldwin, 2016) confirmed empirically that sentence-level embeddings consistently outperform document-level alternatives for short texts, directly justifying the choice of SBERT for the typically brief IVECO ticket texts. The multilingual variant applied in the thesis extends this capability across the German, Italian, and English texts that coexist in the IVECO corpus, and is derived from the XLM-RoBERTa multilingual pre-training lineage of (Conneau et al., 2020).

The thesis employs HDBSCAN (Campello et al., 2013; McInnes et al., 2017) to partition the resulting embedding space: the algorithm discovers clusters of varying density without a pre-specified cluster count and assigns idiosyncratic tickets to a noise class rather than forcing them into inappropriate groups, both properties critical for heterogeneous real-world ticket data. (McInnes et al., 2018) contributed UMAP as a complementary dimensionality reduction step that improves HDBSCAN cluster quality on high-dimensional embeddings. The closest published instantiation of this pipeline is BERTopic (Grootendorst, 2022), which integrates SBERT, UMAP, HDBSCAN, and class-based TF-IDF labeling into a unified system and is the primary reference point from which the thesis differentiates itself. The thesis extends BERTopic through hierarchical vehicle-model partitioning, applying the full pipeline separately to each IVECO model's ticket subset, producing model-specific clusters directly actionable for curriculum planning at the vehicle-platform level, a structural innovation not present in any prior published system.

2.3.2 Strengths and Limitations

The pipeline offers three operational advantages: it produces coherent, reproducible clusters evaluable through established metrics (Newman et al., 2010; Röder et al., 2015); it requires no labeled training data; and it is computationally deterministic, deployable without GPU infrastructure. Its fundamental limitation is that statistical clustering cannot reason about organizational implications. A coherent cluster grouping tickets about AdBlue dosing failures can be identified as semantically consistent, but determining whether it constitutes a training gap requiring a specific course revision, rather than a documentation gap requiring a technical bulletin, or a parts supply issue requiring procurement action, demands domain knowledge and inferential reasoning

entirely outside HDBSCAN's computational scope. The c-TF-IDF keyword labels produced by BERTopic and similar systems are interpretable to domain experts but are not structured, machine-consumable recommendations that a training management system can act on without significant expert interpretation effort. These limitations collectively establish that classical NLP is a necessary but insufficient component of a complete training intelligence system: an LLM reasoning layer is required to bridge the gap between the reproducible cluster structure that NLP produces and the explained, actionable training recommendations that curriculum planners need.

2.4 Large Language Models in Enterprise Text Analytics

2.4.1 LLM Capabilities and Structured Output Generation

(Brown et al., 2020) established that large language models perform diverse NLP tasks through few-shot or zero-shot prompting without task-specific fine-tuning, a capability essential in the IVECO context, where labeled training-gap annotation data for fine-tuning does not exist. (Ouyang et al., 2022) demonstrated that instruction tuning through reinforcement learning from human feedback dramatically improves reliability for structured output tasks, enabling instruction-tuned variants to consistently produce the structured JSON outputs, encoding categorization decisions, gap severity scores, and course recommendations, that the thesis's pipeline requires. (Zhao et al., 2023) confirmed that 7-billion-parameter instruction-tuned models are capable and practically deployable for enterprise analytical tasks, providing direct authorization for the thesis's model choice.

(Wei et al., 2022) demonstrated that chain-of-thought prompting, eliciting step-by-step intermediate reasoning rather than direct answers, dramatically improves LLM performance on multi-step inference tasks and produces outputs that are more accurate, transparent, and auditable for organizational stakeholders. The thesis's four-step prompting strategy, categorize the cluster by technical domain, assess training relevance, evaluate gap severity against retrieved catalogue content, and recommend specific courses, is a direct implementation of chain-of-thought reasoning, transforming the LLM from a text generator into a structured analytical reasoner whose intermediate

steps can be logged and validated by IVECO Academy experts.

2.4.2 Retrieval-Augmented Generation and Local Deployment

(Lewis et al., 2020) introduced Retrieval-Augmented Generation (RAG) as the foundational architecture for grounding LLM outputs in retrieved external documents rather than parametric model memory alone. RAG is architecturally necessary here because, as (Mallen et al., 2023) demonstrated empirically, LLMs are systematically unreliable for domain-specific knowledge absent from pre-training: an LLM prompted to recall IVECO course names from memory would confabulate them entirely, producing recommendations of zero practical utility. The thesis implements a modular RAG architecture, within (Gao et al., 2023)'s taxonomy, in which catalogue retrieval precedes structured chain-of-thought reasoning, a combination not previously evaluated for training gap detection. (Peng et al., 2023) and (Shi et al., 2023) each confirmed that grounding LLM generation in retrieved facts significantly improves factual accuracy, reinforcing RAG's role as the pipeline's primary reliability mechanism.

The IVECO context additionally requires fully local, privacy-preserving deployment: support tickets are sensitive proprietary data that cannot be transmitted to external cloud APIs. (Bommasani et al., 2021) identified data governance as a critical foundation model enterprise risk, while (Dettmers et al., 2023) demonstrated through QLoRA that 7-billion-parameter models can be deployed on standard enterprise hardware via 4-bit quantization, resolving the tension between LLM capability and available infrastructure. The thesis deploys Ollama-managed local quantized inference, ensuring no ticket data leaves the enterprise perimeter. (Edge et al., 2024) extended local RAG toward community-level knowledge summarization, suggesting a future enhancement direction for the catalogue retrieval mechanism.

2.4.3 Strengths and Limitations

Instruction-tuned LLMs within a RAG architecture can interpret cluster summaries, retrieve relevant catalogue entries, and produce structured, explained recommendations in a single inference call, a qualitative advance over what classical NLP can provide for the gap detection task. Against this, (Huang et al., 2023) documented comprehensively that LLMs generate confident but factually incorrect outputs, the stochastic parrot problem identified by (Bender et al., 2021), a failure mode particularly consequential

when training planners act on AI recommendations without independent verification. (Kaddour et al., 2023) catalogued additional challenges: context window limits constraining the volume of ticket and catalogue content processable per inference call, batch inconsistency whereby identical prompts produce moderately different structured outputs across runs, and governance difficulties auditing LLM reasoning chains to satisfy organizational accountability requirements. (Sallam, 2023), examining LLM deployment in healthcare, a domain comparable in stakes and privacy requirements to automotive enterprise analytics, found that LLMs provide genuine value when embedded within mandatory expert oversight protocols, directly validating the thesis's expert validation evaluation design. These limitations collectively require embedding the LLM within a structured pipeline providing RAG grounding, low-temperature deterministic inference, and expert verification of all generated recommendations, precisely the architecture developed in this thesis.

2.5 Hybrid NLP–LLM Architectures

2.5.1 Motivations, Existing Patterns, and Research Gap

The limitations of classical NLP and LLMs are complementary: NLP produces stable, reproducible structural outputs but cannot reason about their organizational implications; LLMs perform sophisticated semantic inference but are unreliable for large-scale document clustering and domain-specific knowledge recall. The natural synthesis assigns each component to the tasks for which it is best suited and passes NLP outputs as grounded inputs to the LLM reasoning stage. (Shen et al., 2023) demonstrated through HuggingGPT that an LLM orchestrating specialized modular AI components consistently outperforms either in isolation, the design pattern directly instantiated in this thesis, where HDBSCAN clusters and the LLM reasons about training implications. (Schick et al., 2023) reinforced this through Toolformer, showing that LLMs augmented with external specialized tools dramatically outperform LLM-only baselines, precisely the logic motivating the use of HDBSCAN cluster outputs and retrieved catalogue entries as LLM inputs rather than asking the LLM to perform those operations directly. (Peng et al., 2023), (Shi et al., 2023), and (Sun et al., 2023) each supplied additional empirical support: the first two confirming that retrieved external grounding consistently improves factual accuracy; the third demonstrating through Think-on-Graph

that iterative LLM reasoning over structured knowledge enables multi-hop inference not achievable through ungrounded prompting, suggesting a future knowledge-graph extension for curriculum navigation.

Two consequential gaps emerge from this literature. First, no published study has applied a hybrid NLP–LLM pipeline to identifying organizational training gaps from support ticket data: HuggingGPT, Toolformer, and Think-on-Graph address multi-modal task completion, knowledge-intensive NLP benchmarks, and knowledge graph question answering respectively, none has been designed or evaluated in an HRD or TNA context, and none has connected ticket cluster discovery to training curriculum recommendation. Second, no study has conducted a direct three-way empirical comparison of a pure NLP approach, a pure LLM approach, and a hybrid pipeline on the same organizational dataset for the same training intelligence task; without such evidence, any claim that hybrid architectures outperform their components in this domain rests on theoretical inference rather than empirical demonstration. These two gaps, the absence of hybrid architectures for automated TNA and the absence of a rigorous three-way comparative evaluation on a real enterprise ticket corpus, together constitute the core empirical contribution of this thesis.

2.5.2 Design Hypothesis

The preceding synthesis yields the following design hypothesis: a hybrid pipeline in which SBERT embeddings and HDBSCAN clustering identify coherent thematic clusters in IVECO ticket data, and in which a locally deployed instruction-tuned LLM augmented by RAG retrieval from the IVECO training catalogue reasons about each cluster's training implications, will outperform both a pure NLP baseline and a pure LLM baseline on three dimensions simultaneously: accuracy of gap detection against an expert-annotated golden set, explainability of reasoning provided to training planners, and integration suitability within the IVECO Academy's existing planning workflow. This hypothesis motivates the three-pipeline experimental design of Chapter 4 and the evaluation criteria of Chapter 5, and provides the falsifiable empirical commitment that distinguishes the thesis from prior theoretical work on hybrid NLP–LLM architectures. The specific architecture, SBERT $\rightarrow$ HDBSCAN with vehicle-model partitioning $\rightarrow$ LLM chain-of-thought reasoning $\rightarrow$ RAG catalogue retrieval $\rightarrow$ structured JSON output, constitutes a novel instantiation of the hybrid NLP–LLM design pattern not previously

implemented or evaluated for training needs analysis in any reported organizational context, and the three-pipeline comparison generates the first empirical evidence base for choosing between architectural approaches in this domain.

2.6 Design Science Research in Information Systems

2.6.1 DSR Principles, Artifact Types, and Evaluation Strategies

Design Science Research provides the epistemological and methodological foundation of this thesis. (Simon, 1996) established the philosophical basis, distinguishing the design sciences, which construct and evaluate artifacts that bring new solutions into existence, from the natural sciences that describe existing phenomena. (March & Smith, 1995) operationalized this in IS through a taxonomy of artifact types, constructs, models, methods, and instantiations, with the thesis's implemented pipeline constituting an instantiation, the most concrete artifact type in their taxonomy. (Hevner et al., 2004) codified the canonical seven-guideline DSR framework, covering artifact design, problem relevance, design evaluation, research contributions, rigor, search process, and communication, to which every aspect of this thesis's methodology is explicitly aligned. (Peffers et al., 2007) operationalized DSR into a six-step process model, problem identification, objective definition, design and development, demonstration, evaluation, and communication, whose phases map directly onto the thesis chapter structure, providing the procedural backbone for the research. (Gregor & Hevner, 2013) contributed the Knowledge Contribution Framework, classifying DSR outputs along solution and problem maturity dimensions and designating novel solutions to known problems as "inventions", the classification that precisely describes this thesis, which designs a new computational architecture for the well-recognized TNA problem. (Venable et al., 2016) provided evaluation design guidance through the FEDS framework: the thesis combines artificial-formative evaluation through quantitative coherence metrics with naturalistic-summative evaluation through expert validation with IVECO Academy staff, following FEDS guidance on rigorous DSR evaluation design. (Wieringa, 2014) further argued that DSR is inherently iterative, building and evaluating in concert rather than in strict linear sequence, a principle reflected in the thesis's incremental pipeline refinement across design cycles.

2.6.2 Why DSR Fits This Thesis

The central research question, how to design and evaluate a hybrid AI framework for identifying training gaps from support ticket data in a secure enterprise environment, is a design problem, not a knowledge problem: it asks what a new artifact should be and how it should be built, not whether a hypothesized variable relationship holds empirically. Alternative paradigms are ill-suited: survey methodology can document TNA inadequacies, as (Ferreira & Abbad, 2013) demonstrated comprehensively, but cannot design the remedial artifact that those inadequacies demand; controlled experimentation can test isolated algorithm performance on benchmark tasks but cannot address how heterogeneous components, NLP clustering, LLM reasoning, RAG retrieval, and expert validation, combine into a deployable organizational system; case study research can describe IVECO's existing TNA practices richly but cannot produce the novel artifact that constitutes the thesis's primary contribution to knowledge. DSR, as (Simon, 1996) argued philosophically and (Hevner et al., 2004) formalized methodologically, is uniquely capable of producing scientifically rigorous knowledge through the design, implementation, and evaluation of innovative IS artifacts that simultaneously advance theoretical understanding and practical organizational capability. That dual advancement, producing a novel artifact that is both theoretically grounded in prior IS and HRD literature and immediately deployable within IVECO's operational context, is the defining characteristic of this thesis's scientific contribution.

2.7 Synthesis and Research Positioning

The six sections of this review collectively map a landscape in which significant theoretical and empirical resources address each component of the research challenge, TNA methodology, ticket analytics, classical NLP, LLM deployment, hybrid architecture design, and DSR methodology, yet in which those resources have never been brought into synthesis for the specific purpose of automated training needs analysis from technical support ticket data. Five literature gaps emerge from this mapping, each representing a distinct dimension along which the present thesis makes a novel contribution; together, they define the complete intellectual space that the thesis occupies.

The first gap concerns TNA methodology. Despite the theoretical robustness of (Goldstein & Ford, 2002)'s three-phase model and the organizational learning infras-

tructure envisioned by (Kraiger, 2014) and (Marsick & Watkins, 2003), (Ferreira & Abbad, 2013) established empirically that TNA remains predominantly reactive and subjective, with no automated data-driven alternatives at operational scale. The thesis addresses this by designing a fully operational AI system that performs continuous, automatic TNA from support ticket data, translating the long-articulated aspiration of analytics-driven TNA into a working artifact, while operationalizing the informal learning signal logic of (Manuti et al., 2015) and (Cerasoli et al., 2018) into a concrete computational pipeline.

The second gap bridges operational ticket analytics and organizational learning. The entire NLP literature on support ticket analysis, surveyed in (Zangari et al., 2023), grounded in (Blei et al., 2003), evaluated through (Röder et al., 2015) and (Newman et al., 2010), and methodologically paralleled by (Rosen & Shihab, 2016), has pursued exclusively operational objectives. No study has connected ticket cluster outputs to training curriculum decisions. The thesis crosses this disciplinary boundary between information retrieval and HRD, establishing for the first time an empirically evaluated connection between automated ticket clustering and training gap identification.

The third gap concerns enterprise LLM deployment under privacy constraints. While (Lewis et al., 2020) and (Mallen et al., 2023) establish the necessity of RAG for domain-specific tasks, and (Dettmers et al., 2023) demonstrates the technical feasibility of local quantized deployment on standard enterprise hardware, empirical evaluations of fully local, privacy-first RAG pipelines on enterprise organizational tasks remain scarce. The thesis contributes a documented design, implementation, and evaluation of an Ollama-based RAG architecture in which no proprietary ticket data leaves the enterprise perimeter, addressing the governance risk that (Bommasani et al., 2021) identified as one of the most significant barriers to enterprise foundation model adoption.

The fourth gap concerns comparative empirical evaluation of NLP–LLM architectural approaches for organizational text analytics. The theoretical case for hybrid designs is well supported by HuggingGPT (Shen et al., 2023), Toolformer (Schick et al., 2023), and chain-of-thought prompting (Wei et al., 2022), but no study has conducted a direct three-way comparison, pure NLP versus pure LLM versus hybrid, on the same organizational corpus for the same training intelligence task. Without such evidence, architectural choices in this domain are driven by theoretical preference rather than empirical data. The thesis's three-pipeline experimental design generates the first such evidence, providing a rigorous empirical basis for the comparative advantage claims

made on behalf of the hybrid architecture.

The fifth gap concerns the application of DSR to systems integrating NLP and LLM pipelines with organizational learning outcomes. AI-augmented HRD artifacts designed and evaluated under the full DSR paradigm as inventions, in (Gregor & Hevner, 2013)'s sense of a novel solution to a known problem, remain exceedingly rare in the published IS literature. The thesis applies the full DSR methodological apparatus to this emerging intersection, contributing both the working artifact and a documented, replicable methodology for artifact design that future researchers can extend.

This thesis addresses all five gaps simultaneously by integrating TNA theory (Ferreira & Abbad, 2013; Goldstein & Ford, 2002; Noe, 2020) with proven classical NLP methods (Campello et al., 2013; Grootendorst, 2022; Reimers & Gurevych, 2019), locally deployed LLM reasoning augmented by retrieval-grounded generation (Lewis et al., 2020; Mallen et al., 2023; Wei et al., 2022), a rigorous three-pipeline empirical comparison, and a Design Science Research methodology (Gregor & Hevner, 2013; Hevner et al., 2004; Peffers et al., 2007), through the design, implementation, and evaluation of a single novel artifact, the IVECO AI skill-gap detection framework. Chapter 3 operationalizes this positioning into a formal research design, specifying the research questions, artifact design objectives, evaluation criteria, and DSR process structure that will govern the remainder of the study.

3 Research Design and Methodology

3.1 Design Science Research Framework

This thesis is grounded in the Design Science Research (DSR) paradigm, an epistemological tradition in Information Systems that focuses on creating and evaluating innovative artifacts designed to address identifiable organisational problems. Unlike behavioural science research, which seeks to explain phenomena through descriptive theory, DSR pursues prescriptive knowledge — asking how systems should be designed to achieve desired outcomes (Hevner et al., 2004). This orientation is appropriate for the present study, which does not seek to explain why training gaps exist but to construct, demonstrate, and evaluate a practical solution for detecting them from unstructured operational data.

(Hevner et al., 2004) situate DSR at the intersection of the environment — the people, organisations, and technologies constituting the problem space — and the knowledge base of existing theories, methods, and artefacts. Here, the environment is IVECO's THD, where support tickets are generated daily but remain analytically dormant with respect to workforce development, and the knowledge base comprises literature on natural language processing, large language models, training needs analysis, and learning analytics.

The procedural execution follows the six-phase process model of (Peffers et al., 2007), beginning with problem identification — the absence of a systematic mechanism for translating ticket content into training intelligence — and proceeding through solution objectives, design, demonstration, evaluation, and communication. Evaluation findings feed back iteratively into design refinements through a human-in-the-loop validation mechanism. Consistent with (Gregor & Hevner, 2013) taxonomy, the KI-Academy framework qualifies as an instantiation — a fully implemented system embodying design knowledge applicable to the broader class of problems involving the transformation of unstructured service data into structured learning intelligence.

3.2 Problem Context and Stakeholder Analysis

3.2.1 IVECO Case Description and Stakeholder Roles

IVECO is a multinational commercial vehicle manufacturer whose internal training unit, IVECO Academy, designs and delivers curricula covering vehicle systems from powertrain diagnostics and electrical fault-finding to telematics and software updates. The Technical Hotline Desk (THD) processes support tickets from dealers and service partners, each documenting a technical problem, diagnostic steps, and resolution — a continuous stream of operational records encoding first-hand evidence of knowledge deficits in the field.

Three principal stakeholder groups shape the artifact's design requirements. THD operators generate the support tickets constituting the system's primary input; their documentation conventions determine the linguistic and structural characteristics of the data. IVECO Academy training planners are the primary intended users of the system's output, deciding which topics warrant new or revised modules and driving requirements for actionable, catalogue-aligned recommendations. Technical subject-matter experts and department managers participate in the human-in-the-loop validation protocol, providing domain knowledge necessary to verify gap assessments. The concerns and information needs of these three groups informed both the functional specification of the artifact and the evaluation methodology.

3.2.2 AS-IS Process Analysis and Identified Pain Points

The AS-IS process, depicted in Figure 3.1, consists of two organisationally separate lanes operating without systematic data exchange. In the THD lane, tickets are created, resolved, and stored in the operational database; once closed, their content remains locked with no mechanism to communicate it to the training function. In the Academy lane, training reviews rely on lagging indicators — expert interviews and annual dealer surveys — rather than systematic analysis of the ticket corpus. As (Ferreira & Abbad, 2013) observe, approaches relying exclusively on expert opinion are prone to anchoring bias and systematic under-identification of distributed knowledge gaps. The consequence is structural misalignment: new technical issues may persist for months before triggering a curriculum response, while a growing and increasingly complex vehicle portfolio makes

expert-based assessment progressively less tractable.

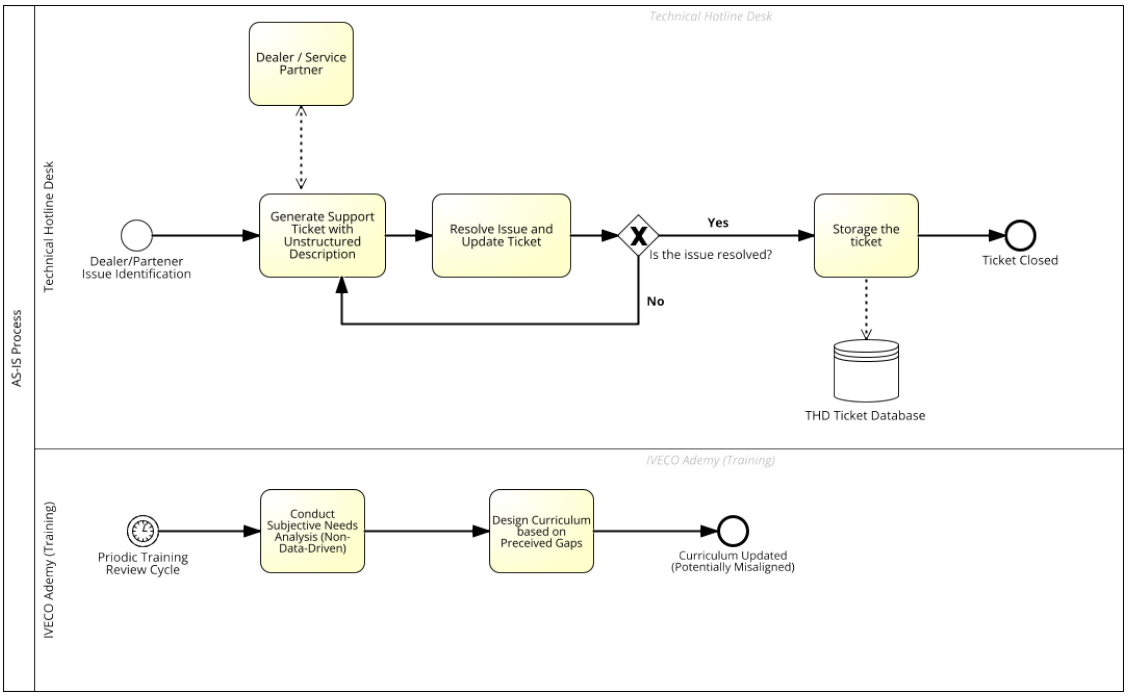


Figure 3.1: Current AS-IS process at IVECO — THD and Training Planning (BPMN)

3.2.3 TO-BE Process Vision and Expected Improvements

The TO-BE process, illustrated in Figure 3.2, introduces a systematic data bridge between the THD ticket database and IVECO Academy's curriculum planning function, replacing episodic expert recall with continuous, evidence-driven analysis. Both the ticket database and the training catalogue serve as direct inputs to the Hybrid AI Framework, which produces a structured skill-gap dashboard that becomes the primary input to the Academy's training review. The expected improvements are threefold: the review process becomes continuous rather than periodic; anchoring decisions in quantitative evidence reduces individual expert bias; and direct catalogue mapping makes every output immediately actionable, giving planners both a curriculum diagnosis and concrete guidance on where development investment is warranted.

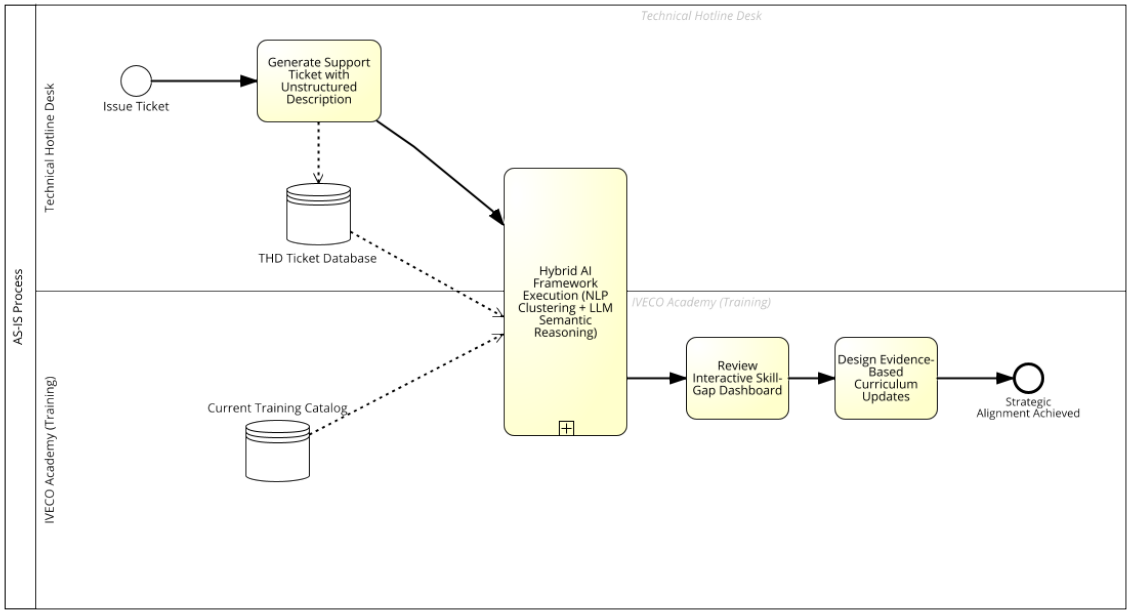


Figure 3.2: TO-BE process enabled by the AI framework — evidence-based curriculum design (BPMN)

3.3 Conceptual Foundation: Training Gap Taxonomy

3.3.1 Definition and Typology of Training Gaps

A training gap, as operationalised in this research, refers to a discrepancy between the knowledge demanded by a recurring technical challenge documented in support ticket data and the training provision available in the organisational curriculum. This definition draws on the TNA framework of (Goldstein & Ford, 2002), which distinguishes organisational, task, and person analysis as three levels of needs identification. The artifact operates at the task level — examining ticket content to identify which technical tasks employees cannot perform independently — and at the organisational level — assessing whether the curriculum adequately addresses those tasks.

The gap typology comprises three categories intrinsic to the artifact's analytical logic. `kein_gap` (no gap) applies when a topic is sufficiently covered by existing courses: the LLM classifies it as fully covered and generates a course recommendation, signalling that no new development investment is required. `inhaltlich_teils_offen` (partially open) applies when a related course exists but does not fully address the specific challenge — a general electrical diagnostics course may only partially cover tickets about a specific

fault code on a particular vehicle model; these topics receive both a recommendation for the closest existing course and a structured proposal for the additional content required. `komplett_neu` (completely new) identifies clusters for which no relevant course exists, triggering the artifact's most detailed output: a module proposal including a suggested title, content outline, competence field, and priority rating reflecting ticket volume. The LLM is explicitly instructed to assign one of the three labels to each topic cluster, with the classification directly determining the downstream output — making the typology a functional design element rather than merely an analytical framework.

3.3.2 Competence Mapping Framework

The artifact also assigns each detected topic to a defined technical competence field — Elektrik, Motor, Telematics, Hydraulics, Braking Systems, ADAS, and others — allowing training planners to group results by technical domain and identify which vehicle system areas are most frequently implicated in support requests. Assignment is performed by the LLM during its reasoning phase, drawing on cluster keywords and example tickets together with its pre-trained understanding of commercial vehicle systems. This semantic approach is preferable to rule-based keyword matching because the same underlying competence deficit frequently surfaces through varied terminology. When exported to IVECO Academy's Power BI environment, competence field classifications become filterable dimensions enabling domain-specific gap summaries, deficit-area prioritisation, and longitudinal curriculum coverage tracking.

3.4 System Requirements

3.4.1 Functional Requirements

The system must ingest a CSV ticket export (`Ticket_ID`, subject, description) and a CSV training catalogue (titles, content, competence area) without requiring technical expertise from end users, then execute a preprocessing pipeline that normalises ticket language and resolves vehicle codes and abbreviations to standardised tokens, followed by dense semantic vector generation and unsupervised clustering that discovers coherent topic groups without a pre-specified cluster count, labelling each with human-

interpretable names and keywords. For each cluster, semantic gap analysis against the catalogue must produce the output appropriate to the gap type: a course recommendation with confidence score for fully covered topics; a nearest-course recommendation plus partial content proposal for partially covered topics; or a complete new module proposal — title, content outline, competence field, and priority rating — for completely new topics. All outputs must be accessible through a non-technical web dashboard and exportable in Power BI-compatible format.

3.4.2 Non-Functional Requirements

Complete data localisation is paramount: all operations — embedding, clustering, and LLM inference — must execute on local hardware with no network egress, since support tickets contain operationally sensitive and personally identifiable information subject to GDPR. The system must be architecturally modular so that NLP models and LLM versions can be replaced without cascading changes, and configurable through a single YAML file without source code modification. The NLP pipeline must complete within approximately thirty minutes on standard enterprise hardware without GPU acceleration, and the system must degrade gracefully under data quality issues, logging LLM schema deviations without aborting the analysis run.

3.4.3 Organisational and Compliance Requirements

GDPR compliance is achieved through data localisation combined with a preprocessing anonymisation step removing personal identifiers before ticket content enters any analytical process, supplemented by role-based access controls. Every gap-type classification must be accompanied by a structured natural-language justification articulating the LLM's reasoning — satisfying the explainability requirement, supporting organisational accountability for AI-informed curriculum decisions, and enabling planners to audit and challenge recommendations. The system must also provide a human validation mechanism through which subject-matter experts can confirm or override individual gap assessments, ensuring the AI augments rather than supplants professional judgment.

3.5 Artifact Design: Three-Pipeline Comparative Framework

3.5.1 Artifact Type and Contribution Claim

Following (Gregor & Hevner, 2013) taxonomy, the KI-Academy framework constitutes an improvement contribution: it applies existing NLP and LLM technologies in a novel configuration to a problem domain — AI-driven training needs analysis from service ticket data — not previously systematically explored. As an instantiation in the DSR sense, it is a fully deployed and functional system rather than an abstract method or conceptual model, embedding design knowledge that can be extracted and applied to analogous organisational settings facing similar challenges in data-driven workforce development. The specific contribution claim is that a hybrid architecture in which classical NLP provides statistically stable topic representations subsequently submitted to LLM semantic reasoning produces training gap analyses of superior quality, explainability, and computational efficiency compared to either a purely classical or a purely LLM-based approach. This claim is evaluated empirically through the three-pipeline comparative framework described in the sections that follow.

3.5.2 High-Level Conceptual Architecture

The artifact comprises four sequential processing layers that transform raw ticket data into structured training-gap intelligence. The ingestion layer receives and validates the ticket export and training catalogue before persisting them as the immutable inputs for a given analysis run. The NLP processing layer applies linguistic normalisation, semantic embedding, and unsupervised clustering to the ticket corpus, producing coherent topic clusters that represent recurring patterns of technical difficulty. The LLM reasoning layer receives these clusters together with the relevant subset of the training catalogue and performs gap-type classification, course recommendation, and module proposal generation for each cluster. The visualisation and export layer presents the complete analysis through a web-based dashboard and provides structured CSV exports formatted for direct import into IVECO Academy's Power BI reporting environment.

The three comparative pipelines share this overarching architecture but differ in how they distribute analytical responsibility between the NLP and LLM components.

Pipeline A performs the full clustering using mathematical methods alone and maps topics to the curriculum via keyword-based similarity computation. Pipeline B delegates both clustering and gap analysis entirely to the language model, bypassing classical NLP. Pipeline C combines both: classical NLP performs clustering to produce statistically stable topic representations, which are then submitted to the LLM for semantic gap analysis augmented by retrieval-augmented generation from the training catalogue. The design rationale for each is developed in the following subsections.

3.5.3 Pipeline A — Classical NLP Approach

Pipeline A establishes the performance ceiling achievable through mathematical text analysis methods alone, without recourse to large language model reasoning. Ticket texts are preprocessed through a series of normalisation steps that include entity normalisation — replacing vehicle model codes, fault code expressions, and technical abbreviations with standardised tokens using regular expression patterns — lemmatisation, and domain-specific stopword removal. The cleaned texts are then transformed into 384-dimensional dense vector representations using a pre-trained multilingual sentence embedding model following the Sentence-BERT architecture introduced by (Reimers & Gurevych, 2019). The selection of a multilingual model is necessary because the IVECO ticket corpus is predominantly German but frequently incorporates technical terms drawn from Italian and English.

Clustering is performed using HDBSCAN, the hierarchical density-based spatial clustering algorithm formalised by (Campello et al., 2013). HDBSCAN was selected over alternatives such as K-Means or LDA topic modelling for three reasons directly relevant to the ticket clustering problem. First, it does not require the number of clusters to be specified in advance, which is essential given that the number of distinct technical topics in the corpus is neither known a priori nor fixed across analytical periods. Second, it handles non-uniform cluster densities and varying cluster shapes characteristic of a heterogeneous ticket corpus where some issues generate dense concentrations of similar tickets while others produce sparse, isolated cases. Third, it assigns a noise label to tickets that belong to no coherent cluster rather than forcing them into artificial groupings — an honest signal about low-frequency idiosyncratic issues. Following clustering, Pipeline A extracts representative keywords per topic using c-TF-IDF weighting and constructs labels following a vehicle–system–symptom naming

convention. For curriculum mapping, it computes cosine similarity between topic keyword representations and catalogue content fields. The primary limitation is that this similarity-based mapping captures only lexical overlap and cannot perform the conceptual reasoning required to detect partial coverage gaps or propose new content.

3.5.4 Pipeline B — LLM-Only Approach

Pipeline B assesses the extent to which a large language model can perform both clustering and gap analysis when applied directly to raw ticket texts without prior NLP preprocessing. Batches of raw tickets are submitted to the locally deployed model with a prompt instructing it to identify thematic clusters, assign descriptive labels, and evaluate coverage against the training catalogue. Batching is necessitated by the context window limitations of the local model. While Pipeline B's appeal lies in its simplicity — eliminating the NLP layer reduces the number of components and configuration decisions — it introduces critical failure modes at operational scale. The most significant is the lack of global context across batches: clusters identified in one batch may be conceptually equivalent to clusters identified in a different batch under a different label, producing an inconsistent and fragmented topic structure. Additionally, the per-token computational cost of running inference over raw ticket text is substantially higher than the vector mathematics employed by the NLP pipeline, making Pipeline B considerably slower for large datasets. Pipeline B is included in the comparative framework precisely because its limitations provide the empirical contrast required to justify the additional complexity of the hybrid Pipeline C design.

3.5.5 Pipeline C — Hybrid NLP → LLM Approach

Pipeline C, the hybrid approach, is the primary artifact of this research and embodies the design hypothesis that combining classical NLP with LLM reasoning yields superior results to either approach in isolation. The pipeline's fundamental architectural insight is that the two technologies are genuinely complementary: classical NLP excels at the statistical aggregation of large corpora into compact and terminologically stable representations but lacks semantic understanding; LLMs excel at contextual reasoning and nuanced interpretation but struggle with the scale, consistency, and computational cost demands of processing thousands of raw documents directly. Pipeline C exploits this complementarity by assigning each technology to the sub-task it performs best,

producing a design that is stronger than either component individually.

The NLP layer executes the full clustering analysis described for Pipeline A, producing topic clusters each represented as a compact analytical package — a topic identifier, descriptive label, ranked keyword list, and three representative example tickets. This compact representation dramatically reduces the token cost of subsequent LLM analysis compared to Pipeline B, because the model receives a distilled summary of each topic rather than the full text of all associated tickets. The LLM reasoning phase is further enhanced by a Retrieval-Augmented Generation (RAG) module following the architectural pattern of (Lewis et al., 2020). Rather than embedding the full training catalogue in the system prompt — which would rapidly exhaust the model's context window — the RAG module computes cosine similarity between a query embedding constructed from the topic's label and keywords and the vector representations of all catalogue entries, injecting the top-ranked courses into the system prompt as grounded informational context for the gap assessment. This retrieval-then-reason architecture ensures the LLM classifies gaps against actual catalogue content rather than relying on parametric knowledge alone, combining NLP-derived topic stability, RAG-enhanced catalogue awareness, and LLM semantic reasoning to satisfy both the analytical accuracy and operational practicality requirements identified in the system design.

3.6 Evaluation Methodology

3.6.1 Quantitative Effectiveness Metrics

The quantitative evaluation applies identical datasets, metrics, and infrastructure to all three pipelines. The primary metric is gap detection accuracy — the proportion of gap-type classifications agreeing with expert-annotated ground truth — computed per category and as a weighted aggregate. Further metrics are topic coherence (C_V score), coverage ratio, end-to-end processing efficiency, and curriculum mapping precision. All values are reported in Chapter 5.

3.6.2 Qualitative Integration Suitability Criteria

The qualitative evaluation follows the naturalistic evaluation framework of (Venable et al., 2016) and assesses four integration suitability criteria through the expert polling instrument, given that technical performance alone is insufficient for sustained organisa-

tional adoption (Wang et al., 2019). Output interpretability addresses whether outputs are clear and actionable for training planners with limited AI familiarity. Explainability of reasoning examines whether gap classification justifications are logically coherent — particularly salient when contrasting Pipeline A's computationally derived rationale with Pipeline C's LLM-generated prose. Perceived relevance to practice captures whether expert participants consider recommendations aligned with their own professional judgment. Integration feasibility reflects their assessment of whether the pipeline could realistically be adopted into IVECO Academy's workflow without prohibitive effort.

3.6.3 Expert Annotation and Validation Protocol

Expert validation is conducted through a structured online polling instrument distributed to internal IVECO Academy trainers, hosted through Hochschule Neu-Ulm's platform. Five sections cover the major evaluation dimensions. Section one assesses NLP clustering quality via bubble-chart and card-based cluster views rated on a five-point Likert scale. Section two evaluates the LLM pipeline's three output sub-views — the Skill Gaps table (gap type, competence field, relevance, and justification), the Course Recommendations table (recommended course, cosine-similarity trust score, and rationale), and the Proposed New Trainings table (title, area, and content outline) — each rated separately. Section three rates Flask dashboard usability; section four assesses the overall methodological approach; section five captures adoption likelihood and collects open-text improvement suggestions. Quantitative ratings are analysed descriptively; open-text responses undergo qualitative content analysis. Full results are reported in Chapter 6.

3.6.4 Evaluation Dataset Construction and Threats to Validity

The evaluation dataset consists of anonymised THD support tickets divided into an analytical corpus processed by all three pipelines and a held-out validation subset for expert annotation. Ground truth labels are established through independent expert annotation; inter-rater agreement is assessed using Cohen's kappa and disagreements resolved through structured reconciliation.

Three categories of threats to validity are acknowledged. Internal validity is threatened by annotation subjectivity — expert labels may reflect individual interpretive

conventions introducing bias — mitigated by the multi-annotator design and kappa statistics. External validity is constrained by the single-organisation, single-industry case, limiting transferability of specific findings, though the embedded design knowledge — the three-pipeline architecture, gap typology, RAG-enhanced mapping, and validation protocol — is intended to generalise. Construct validity concerns centre on the gap taxonomy's three-category simplification of a continuous competence landscape; definitional guidance in both the LLM prompt and annotation protocol reduces boundary ambiguity, and disagreement patterns are examined qualitatively to inform future refinement.

4 System Architecture and Implementation

This chapter describes how the conceptual design established in Chapter 3 was realised as a functioning software artifact. It proceeds from the overall system design through each of the three comparative pipelines, the Retrieval-Augmented Generation module, the user interface, and the implementation quality measures. Design decisions are grounded in the requirements of Chapter 3 and traceable to the evaluation criteria assessed in Chapter 5.

4.1 Overall System Design

4.1.1 Architectural Overview and Technology Stack

The KI-Academy framework operates as a local, privacy-first web application. Operating entirely on-premise was mandated by the compliance constraints established in Chapter 3: IVECO support ticket data is operationally sensitive, making cloud transmission inadmissible. The application layer uses Flask to expose a web interface while orchestrating background analysis tasks, with Jinja2 as the server-side template engine and visual styling following IVECO Academy brand guidelines.

The NLP subsystem uses the `paraphrase-multilingual-MiniLM-L12-v2` model from the `sentence-transformers` library to encode ticket texts as 384-dimensional dense vectors. This Sentence-BERT model (Reimers & Gurevych, 2019) handles German-language technical vocabulary with fidelity comparable to its English training distribution, a prerequisite for the predominantly German THD ticket corpus, while remaining light enough for CPU-only inference on standard enterprise hardware. Clustering is performed by HDBSCAN (Campello et al., 2013; McInnes et al., 2017) and additional NLP operations, TF-IDF keyword extraction and cosine-similarity retrieval, by `scikit-learn`.

The LLM subsystem uses Ollama to serve `qwen2.5:7b-instruct` locally; the instruction-tuned variant was chosen because it reliably follows structured formatting directives and produces consistent JSON outputs essential for downstream parsing

of gap-analysis results. Inference runs at temperature 0.2 and a 4,096-token context length, with the low temperature biasing outputs toward highest-probability, factually grounded completions. All parameters, paths, model identifiers, clustering granularity, and RAG retrieval depth, are centralised in a `config.yaml` file, enabling reconfiguration without code changes and fully reproducible experimental comparisons across the three pipelines.

4.1.2 Data Flow, Processing Layers, and Component Interaction

The system's data flow is organised into four sequential layers, depicted in Figure 4.2. The ingestion layer accepts two CSV uploads: a ticket export with `Ticket_ID`, `Betreff`, and `Beschreibung` fields, and a training catalogue with `Titel`, `Inhalte`, and `Bereich` fields. Both are validated and stored under `data/current_input` as immutable inputs for a given run. The system adopts a destructive-overwrite strategy for successive runs, ensuring the dashboard always reflects the most recent analysis without stale artefacts contaminating results. This design directly addresses the organisational gap identified in Chapter 3, where the THD ticket database and the training catalogue existed in disconnected lanes with no mechanism for relating one to the other; by accepting both as inputs to a single session, the system creates an integrated view of operational demand and training supply.

The NLP layer produces three structured artefacts, `topics.csv`, `topic_examples.csv`, and `ticket_topic_map.csv`, persisted under `data/processed/nlp_output`, which serve as the handoff interface between the NLP and LLM components. This design feature decouples the two subsystems and enables their independent evaluation across the three pipeline conditions. The LLM reasoning layer consumes these artefacts together with the catalogue and produces `skill_gaps.csv`, `training_recommendations.csv`, and `proposed_new_trainings.csv` under `data/processed/llm_output`. The visualisation and export layer then presents results through the Flask dashboard and packages all artefacts into a ZIP archive for Power BI integration. This layered architecture ensures that raw ticket data flows through normalisation, embedding, clustering, LLM reasoning, and presentation in a directed, auditable sequence. Figure 4.3 provides a comparative overview of how all three pipelines traverse these layers.

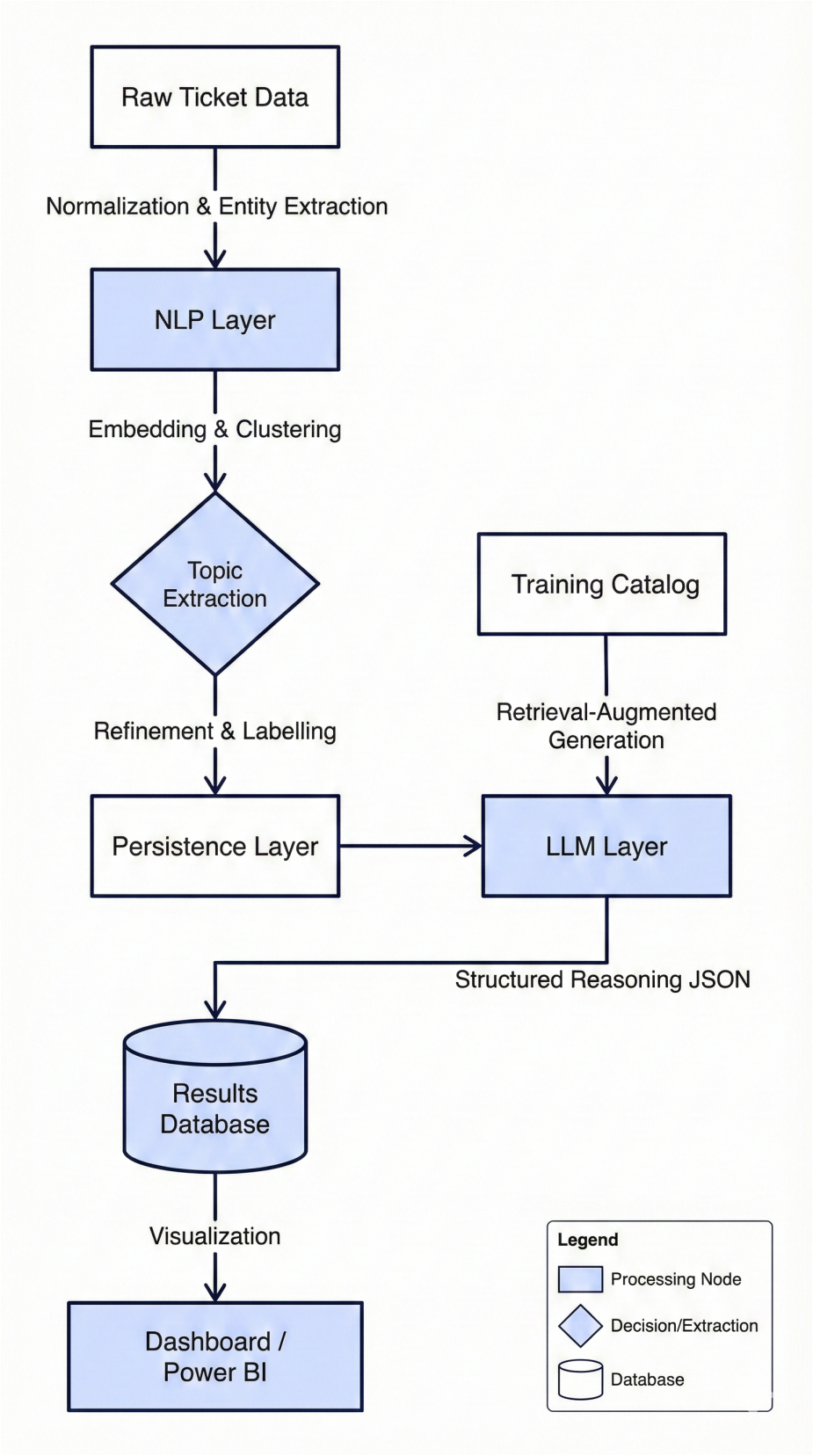


Figure 4.1: High-level system architecture of the KI-Academy framework

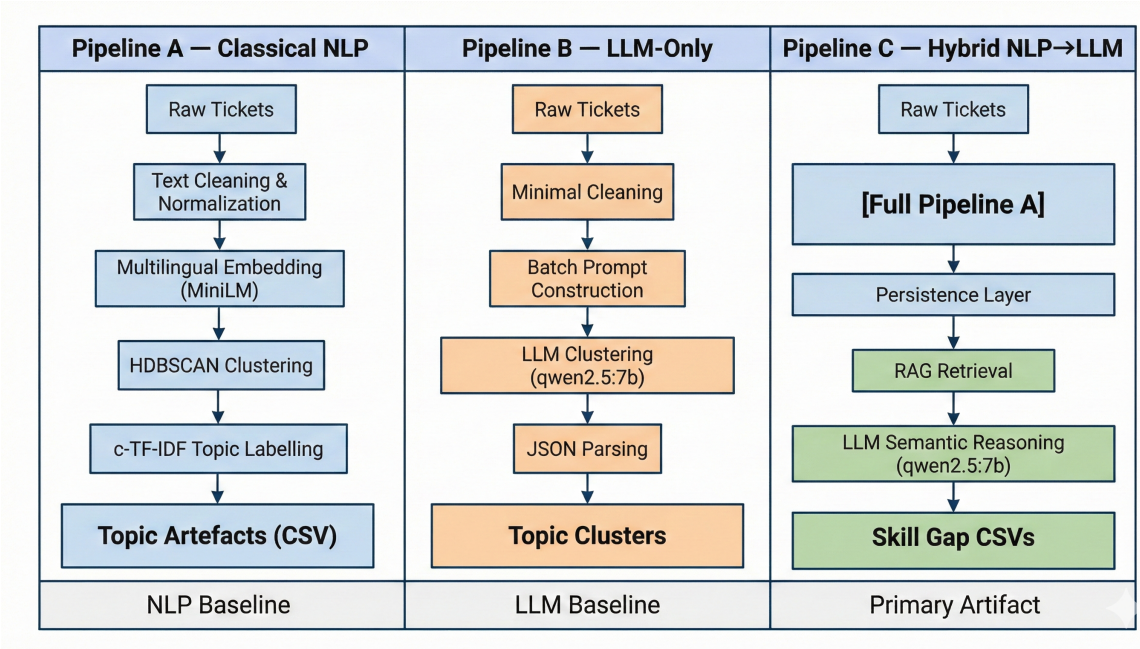


Figure 4.2: Data flow and processing layers of the KI-Academy framework

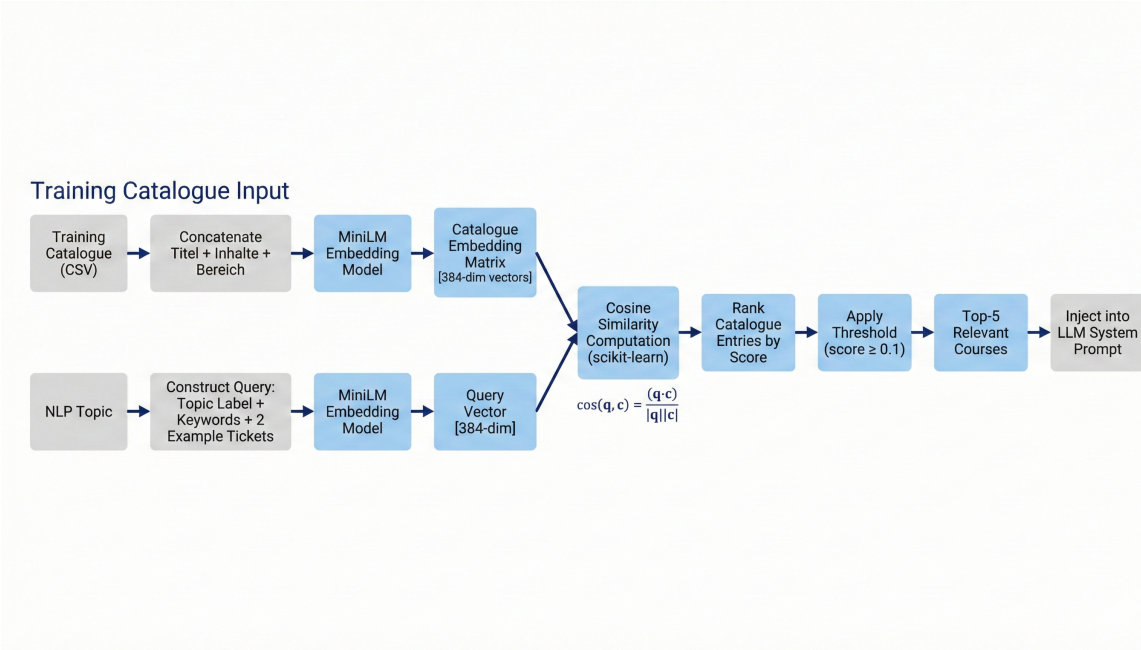


Figure 4.3: Comparative overview of the three pipeline conditions, Pipeline A (NLP), Pipeline B (LLM), and Pipeline C (Hybrid NLP → LLM)

4.2 Pipeline A, Classical NLP Implementation

4.2.1 Preprocessing, Normalization, and Embedding

Pipeline A is the classical NLP baseline, transforming unstructured ticket text into coherent topic clusters without any LLM involvement. Text cleaning normalises to lowercase, retains German-specific characters (ä, ö, ü, ß), removes special characters and whitespace, and concatenates `Betreff` and `Beschreibung` into a single `Full_Text` field, their combination yields a richer representation, as the subject alone often lacks technical context while the description sometimes omits the specific component referenced in the subject. Entity normalisation then applies curated regular-expression patterns to standardise variant spellings of vehicle-model names: "S-Way", "S Way", and "sway" are all mapped to the canonical token `VEHICLE_SWAY`, preventing the multilingual model from treating semantically identical references as distinct entities and thereby fragmenting clusters that should be coherent. The normalised texts are encoded by the MiniLM model into 384-dimensional dense vectors; (Reimers & Gurevych, 2019) demonstrate that Sentence-BERT embeddings substantially outperform word-averaging approaches on similarity tasks, making them well-suited to density-based clustering.

4.2.2 HDBSCAN Clustering and Topic Representation

HDBSCAN was selected over k-means and standard DBSCAN because it requires no a priori cluster count, discovers clusters of varying density, and models noise as a distinct class, properties essential when the number of distinct technical issues in the corpus is unknown and some tickets may not fit any coherent theme. (Campello et al., 2013) demonstrate that HDBSCAN outperforms competing density-based algorithms on benchmark datasets with irregular cluster shapes, which reflects the realistic structure of a technical support corpus. The implementation adds a domain-specific first stage: each ticket is assigned to a vehicle-model bucket based on entity tokens detected during preprocessing, ensuring that tickets about the S-Way and the Daily are never co-clustered even when they share symptom vocabulary, taking advantage of the domain knowledge that technician skill gaps are inherently vehicle-specific. HDBSCAN then runs independently within each bucket, with `min_cluster_size` computed dynamically from the user-selected granularity setting (range: 3–20) and `min_samples` fixed at two

to capture rare but technically significant failure patterns without inflating the noise class. Topic labels are generated via class-based TF-IDF (c-TF-IDF), as in BERTopic (Grootendorst, 2022), weighting terms by within-cluster specificity rather than corpus-wide frequency. Each topic receives a human-readable label in the format *Vehicle, System, Problem*, for example "S-Way, Motor, Injektor defekt", with a domain-specific stopword list filtering generic support-ticket vocabulary.

4.2.3 Gap Detection Logic

Gap detection in Pipeline A operates through the three persistence artefacts rather than autonomous reasoning. The three output files encode the complete topic structure of the ticket corpus: which topics exist, which tickets belong to each, and what the most representative ticket texts are. The `topic_examples.csv` file is particularly significant for evaluation purposes, as it provides human-readable content that a domain expert can inspect to assess whether a given cluster reflects a genuine technical skill gap or merely a procedural or administrative pattern unrelated to training needs. In the comparative evaluation, Pipeline A's gap detection capability is assessed by examining whether its NLP-generated topic labels correspond to areas identifiable as uncovered by the catalogue through expert annotation. Crucially, Pipeline A does not autonomously reason about catalogue coverage, that capability is introduced in Pipeline C. Its purpose as a standalone evaluation condition is to isolate the NLP layer's contribution and establish the coherence baseline against which the LLM-enhanced pipelines are compared.

4.3 Pipeline B, LLM-Only Implementation

4.3.1 Minimal Preprocessing and LLM-Based Clustering

Pipeline B is the LLM-only baseline, designed as a controlled condition to isolate what is lost when the classical NLP layer is removed. Preprocessing is minimal, ticket texts are cleaned of gross formatting noise but neither entity-normalised nor embedded. Concatenated texts are assembled into prompt payloads and submitted directly to Ollama. Because `qwen2.5:7b-instruct`'s 4,096-token context limit precludes processing the full corpus in a single prompt, Pipeline B operates in batches of approximately 50–100 tickets, dynamically sized to remain within the token budget. The model is instructed

to identify 3–8 coherent technical themes per batch and return a structured JSON mapping cluster names to ticket identifiers, following the few-shot instruction-following paradigm of (Brown et al., 2020). No catalogue information is injected at this stage, meaning that any thematic groupings reflect the model's intrinsic understanding of the ticket content rather than a comparison against available training provision.

4.3.2 LLM-Based Gap Reasoning and Limitations

Pipeline B exhibits several critical failure modes at realistic corpus scale. Batch-wise processing introduces inconsistency: the model has no memory of labels assigned in prior batches, so functionally identical topics receive different names across batches, "Motor" in one, "Triebwerk" in another, with no mechanism for post-hoc merging. The model also tends to produce overly broad catch-all clusters when inputs exceed its effective attention span, a manifestation of the "Lost in the Middle" effect documented by (Liu et al., 2024), whereby LLMs assign reduced attention to content in the middle of long context windows. Most critically, because the training catalogue is never injected into the context, gap reasoning rests entirely on the model's parametric memory, which cannot be assumed to contain reliable knowledge of IVECO's specific course offerings. (Mallen et al., 2023) demonstrate that language models are unreliable for precisely this kind of domain-specific, low-frequency knowledge, providing strong grounds for treating Pipeline B as a controlled baseline rather than a viable production approach. These limitations collectively motivate the hybrid design of Pipeline C, in which the NLP layer resolves the clustering problem before the LLM is invoked.

4.4 Pipeline C, Hybrid NLP → LLM Implementation

4.4.1 Stage 1, NLP Topic Extraction and Persistence Layer

Pipeline C, the primary artifact of this thesis, integrates the two preceding approaches into a two-stage architecture that captures the strengths of each while mitigating their respective weaknesses. Stage 1 executes the complete NLP pipeline of Section 4.2, tickets are cleaned, entity-normalised, embedded, clustered using HDBSCAN, and labelled via c-TF-IDF. Its defining design feature is not the NLP processing itself but the persistence

layer that follows: all topic artefacts are committed to the file system in stable, structured form before any LLM invocation occurs. This persistence-first approach serves two purposes: it decouples the NLP and LLM stages, each can be executed, restarted in case of failure, and evaluated independently as separate experimental conditions, and it provides Stage 2 with a compact, structured summary of the corpus rather than raw texts, circumventing the context-window limitations that afflict Pipeline B. Each topic is summarised by its label, top ten c-TF-IDF keywords, three centroid-proximal representative tickets, and aggregate ticket count, a representation typically an order of magnitude smaller in token volume than the raw texts it distils, allowing the LLM to reason over the full topic space of a large corpus within a series of individually context-bounded prompts.

4.4.2 Stage 2, LLM Semantic Reasoning and Output Integration

Stage 2 subjects each persisted NLP topic to structured semantic reasoning by `qwen2.5:7b-instruct`. The model receives a system prompt establishing its expert persona and injecting the top-five catalogue entries retrieved by the RAG module (Section 4.5), followed by a user prompt with the topic's label, keywords, representative examples, and ticket count. It is instructed to reason through four sequential steps: assigning a Kompetenzbereich (e.g., Elektrik, Motor, Telematik); rating training relevance as High, Medium, or Low; assessing catalogue coverage as fully covered, partially covered, or absent; and either recommending an existing course or proposing a new training module with title, content outline, and priority rating.

This four-step structure is a direct implementation of chain-of-thought prompting as formalised by (Wei et al., 2022), who demonstrate that decomposing complex reasoning tasks into explicit intermediate steps substantially improves LLM output accuracy and consistency. By requiring competence classification before gap analysis, the prompt design prevents the model from short-cutting to a recommendation without engaging with the topic's technical content. Output is constrained to a predefined JSON schema, fields for `gap_typ`, `begruendung`, `empfohlene_trainings` with trust scores, and `vorgeschlagene_neue_trainings`, reducing the surface area for plausible-but-incorrect free-form statements that cannot be automatically detected. Responses are assembled into the three output CSVs while full raw JSON logs are written to an

audit JSONL file for post-hoc inspection. This separation of statistical clustering in Stage 1 from semantic interpretation in Stage 2 reflects the core design hypothesis of the thesis: that neither approach alone is sufficient, but their combination produces results qualitatively superior to either baseline.

4.5 Retrieval-Augmented Generation Module

4.5.1 Training Catalog Indexing and Retrieval Strategy

The RAG module grounds the LLM's gap analysis in IVECO's actual training catalogue rather than parametric memory. (Lewis et al., 2020) demonstrate that grounding LLM outputs in retrieved documents significantly improves factual accuracy on knowledge-intensive tasks. At startup, all catalogue entries are loaded, their `Titel`, `Inhalte`, and `Bereich` fields concatenated, and each encoded by the MiniLM model into a 384-dimensional vector cached in memory. Using the same embedding model for both ticket clustering and catalogue retrieval ensures that tickets and courses share a common semantic space, maximising the likelihood that topically similar content occupies nearby regions.

For each NLP topic, the RAG module builds a retrieval query by concatenating the topic's label, its top keywords, and the text of its two most representative example tickets. This compound query is embedded and compared against all catalogue vectors using cosine similarity via `scikit-learn`. Cosine similarity is the appropriate metric for normalised dense vectors produced by sentence-transformer models, capturing directional similarity independently of vector magnitude. The top five results are returned, subject to a minimum score threshold of 0.1 to suppress insufficiently relevant entries and prevent the injection of unrelated catalogue material into the LLM context. Five was chosen to balance context richness against the 4,096-token prompt budget.

4.5.2 Context Injection and Prompt Engineering

The five retrieved catalogue entries are injected into the LLM's system prompt as a delimited block labelled `TRAININGSKATALOG`, with an explicit directive that these entries represent the current IVECO Academy catalogue and constitute the sole reference for determining topic coverage. This grounding instruction prevents the model from drawing on parametric knowledge of generic commercial-vehicle training programmes

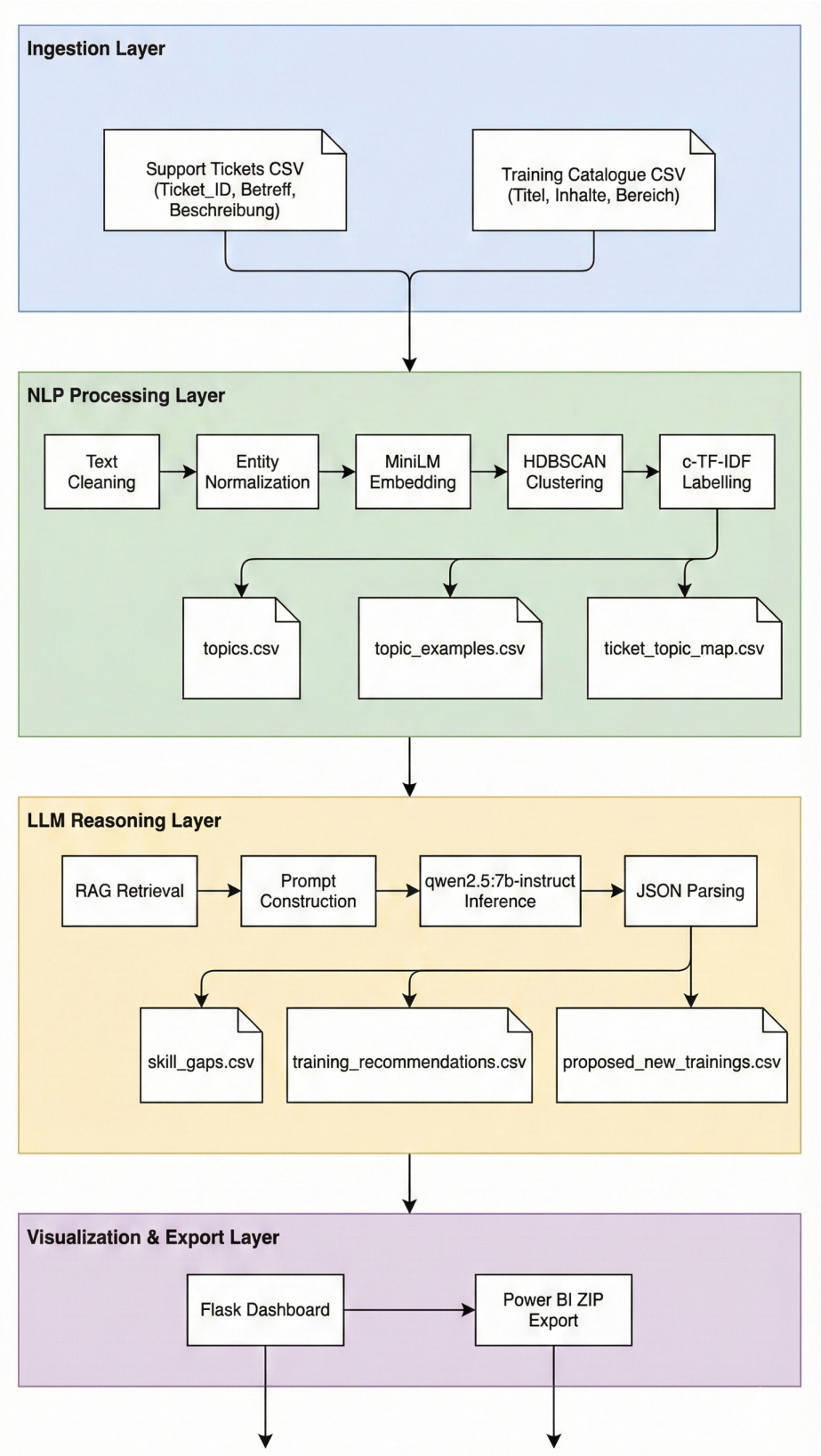


Figure 4.4: RAG module, dual-track embedding and cosine similarity

that may not exist in IVECO's actual catalogue. The per-topic user prompt is kept minimal, topic label, keywords, truncated representative examples, and ticket count, leaving the system prompt to bear the primary context-setting burden. This separation of a richly specified system prompt from a lean, stimulus-specific user prompt follows prompt engineering practices validated by (Wei et al., 2022) for eliciting consistent structured reasoning across inference calls. Full raw JSON responses are logged to an audit JSONL file for post-hoc debugging and evaluation without re-running inference.

4.6 User Interface and Visualization Layer

4.6.1 Flask Web Application and Dashboard Components

The user-facing interface is a Flask web application rendered through Jinja2 templates, designed for non-technical training planners and academy staff. It presents only analysis outputs - topics, gaps, recommendations - without exposing pipeline configuration or internal states. Styling follows IVECO Academy's corporate colour scheme with the primary blue (`#1554ff`) for interactive elements. The main dashboard shows ticket count, course count, and topic count as summary cards alongside a single "Start Analysis" button that initiates the full pipeline as a background task. A screenshot is provided in Appendix A. The data ingestion view renders both uploaded CSVs as paginated previews, surfacing formatting errors before analysis begins.

The NLP results dashboard presents clustering output as an interactive bubble chart - bubble area proportional to ticket count, hover revealing label and top keywords - supplemented by a card grid with drill-down links to individual ticket lists for qualitative coherence validation (Appendix B). The LLM analysis view uses a three-panel tabbed interface: the "Skill Gaps" panel lists each topic with its gap type - Covered, Partially Covered, or Completely New - and the model's justification; the "Recommendations" panel lists existing courses with cosine-similarity trust scores; and the "New Trainings" panel presents proposed modules - titles, content outlines, and priority ratings - for topics absent from the catalogue. A screenshot of this view is included in Appendix C.

4.6.2 Interactive Visualization and Power BI Export

The export route bundles all structured CSV artefacts, `topics.csv`, `topic_examples.csv`, `ticket_topic_map.csv`, `skill_gaps.csv`, `training_recommendations.csv`, and

`proposed_new_trainings.csv`, into a ZIP archive formatted for direct import into Power BI, enabling training planners to build custom visualisations and cross-filter results by vehicle model or competence area within their existing reporting environment. Building a native analytics layer was deliberately out of scope: the framework's contribution lies in the AI-driven analysis pipeline, not in reporting infrastructure IVECO Academy already possesses.

4.7 Implementation Quality Assurance

The implementation is organised around three quality properties: modularity, reproducibility, and privacy compliance. Modularity is achieved by organising the application, NLP pipeline, LLM pipeline, and ingestion module as independent Python packages communicating exclusively through the persistence-layer files, so any component, embedding model or LLM, can be swapped by updating a single entry in `config.yaml` without touching downstream code. Reproducibility is supported by that same `config.yaml`, which version-controls all hyperparameters and model identifiers, and by the JSONL audit log of raw LLM responses, enabling post-hoc re-parsing without re-running inference. Privacy is enforced architecturally: both the MiniLM model and `qwen2.5:7b-instruct` run entirely on local hardware via Ollama, transmitting no ticket data externally, in accordance with the data-localisation requirement established in Chapter 3. Malformed LLM responses are caught and flagged in the output CSV without interrupting the analysis run, ensuring graceful degradation. The following chapter presents the empirical evaluation of all three pipelines against the success criteria defined in the research design.

5 Results

Pipeline C represents the integrated artifact combining the NLP clustering stage (Pipeline A) with the LLM-based gap analysis stage (Pipeline B). The results presented in this chapter report the outputs of each stage sequentially before presenting the integrated end-to-end outcome. Together, these results demonstrate how the framework processes raw service ticket data into actionable, curriculum-aligned training recommendations for the IVECO Academy.

5.1 Dataset Characteristics

The empirical basis for this study comprises 10,305 service tickets drawn from the IVECO Technical Hotline Desk (THD), a live operational support channel through which field technicians report vehicle faults, diagnostic ambiguities, and unresolved repair issues. All tickets were written in German and were anonymized prior to analysis in accordance with organizational data governance requirements. The corpus spans multiple IVECO vehicle lines, including the Daily, S-Way, S-eWay, Eurocargo, Trakker, X-Way, T-Way, and Magirus, thereby capturing the full breadth of technical domains represented in the product portfolio. The training catalogue against which detected gaps were evaluated consists of 185 existing courses maintained by the IVECO Academy. This catalogue served as the retrieval corpus for the RAG-based matching module in Pipeline B, enabling gap classifications to be grounded in the actual availability of training provision. No manual annotation was performed on the ticket corpus prior to pipeline execution; the system was designed to operate unsupervisedly, with expert validation reserved for the subsequent evaluation phase reported in Chapter 6.

5.2 Quantitative Effectiveness Results

5.2.1 Gap Detection Accuracy and Type Classification

The LLM-based gap analysis stage produced classifications for all 15 topics supplied by the NLP clustering stage, including the 14 named thematic clusters and the noise cluster designated *Andere Themen*. Of the 15 topics, 11 were classified as *kein_gap*,

indicating full coverage by existing training provision within the 185-course catalogue. This represents 73.3 percent of all evaluated topics. Two topics were classified as *komplett_neu*, denoting the complete absence of corresponding training material: *Iveco S-eWay – Abgasnachbehandlung (EATS) – Eway* and *Andere Themen*. These topics collectively account for 13.3 percent of the classified set. A further two topics were classified as *inhaltlich_teils_offen*, meaning partial coverage with identifiable content gaps: *Iveco Daily – Fahrwerk / Bremse – Erfahrung* and *Fahrwerk / Bremse – Magirus*, also representing 13.3 percent of the total. Across all 15 topics, 13 were assigned a training relevance rating of *hoch* (high), underscoring the operational significance of the ticket content as a signal for skill development needs. The pipeline generated 14 course recommendations drawn from the existing catalogue, matched across 11 topics, while three topics received no course match, an outcome confirmed as reflecting genuine curricular gaps rather than retrieval failure, consistent with their *komplett_neu* and *inhaltlich_teils_offen* classifications. In addition, four new training proposals were generated: two assigned high priority and two medium priority, providing actionable input for curriculum planning.

5.2.2 Topic Coherence, Coverage Mapping, and Processing Efficiency

Topic coherence was evaluated using the C_V metric introduced by (Röder et al., 2015), which provides a probabilistic measure of semantic relatedness among the top keywords associated with each discovered topic. As implemented within the BERTopic framework (Grootendorst, 2022), coherence scoring draws on word co-occurrence statistics to assess whether the terms defining each cluster represent semantically unified themes. However, given the domain-specific and multilingual nature of the IVECO ticket corpus, C_V scores are most meaningfully interpreted in conjunction with expert validation; accordingly, topic coherence evaluation is reported as part of the expert validation in Chapter 6.

In terms of coverage, the NLP clustering stage successfully assigned 9,276 of 10,305 tickets to named thematic clusters, yielding an overall coverage ratio of 90.0 percent. The remaining 1,029 tickets were allocated to the noise cluster *Andere Themen*, representing 10.0 percent of the total corpus. This noise allocation is consistent with the behaviour of density-based clustering algorithms operating on heterogeneous text collections, where

tickets that do not achieve sufficient local density to form stable clusters are excluded from named groupings rather than being forced into ill-fitting categories.

Processing efficiency was measured on consumer-grade hardware: a Huawei MateBook 14s equipped with an Intel Core i7-11370H processor running at 3.30 GHz and 16 GB of RAM, without GPU acceleration. Pipeline A, encompassing SBERT embedding generation and HDBSCAN clustering for all 10,305 tickets, completed in approximately 11 minutes. Pipeline B, comprising RAG-based retrieval and LLM-driven gap classification for 15 topics, required approximately 17 minutes. The total end-to-end processing time for Pipeline C was therefore approximately 28 minutes. It should be noted that processing time is hardware-dependent and would be significantly faster on dedicated server infrastructure, making the framework operationally viable for periodic re-execution as ticket volumes grow.

5.3 Comparative Pipeline Analysis

5.3.1 Pipeline A Results

Pipeline A constitutes the NLP clustering stage of the system. Operating on the full corpus of 10,305 German-language service tickets, it produced 14 named thematic topic clusters in addition to a noise cluster. Embeddings were generated using the paraphrase-multilingual-MiniLM-L12-v2 model from (Reimers & Gurevych, 2019), capturing semantic similarity across multilingual technical text. Cluster boundaries were determined through HDBSCAN (Campello et al., 2013), a hierarchical density-based algorithm well-suited to discovering clusters of variable size and shape without requiring the number of clusters to be specified in advance.

The 14 named clusters span three dominant technical domains: drivetrain and motor issues, electrical and electronics systems, and aftertreatment and EATS systems. The largest cluster, *Iveco Daily – Fahrwerk / Bremse – Erfahrung*, contains 1,015 tickets, while the smallest named cluster, *Elektrik / Elektronik – Massif*, contains 317 tickets. The full set of clusters and their ticket volumes, in descending order, comprises: *Iveco Daily – Fahrwerk / Bremse – Erfahrung* (1,015), *Iveco Eurocargo – Elektrik / Elektronik – Telematics* (905), *Iveco Bus – Motor / Antrieb – Ärger* (846), *Iveco Trakker – Motor / Antrieb – Eurotrakker* (819), *Iveco S-Way – Elektrik / Elektronik – Way* (800), *Iveco S-eWay – Abgasnachbehandlung EATS – Eway* (768), *Iveco X-Way – Aufbau / Karosserie*

– *Way* (761), *Iveco T-Way – Abgasnachbehandlung EATS – Way* (743), *Motor / Antrieb – Magirus* (532), *Elektrik / Elektronik – Update* (481), *Iveco Daily – Elektrik / Elektronik – Update* (425), *Motor / Antrieb – Massif* (455), *Fahrwerk / Bremse – Magirus* (409), and *Elektrik / Elektronik – Massif* (317). The noise cluster, *Andere Themen*, contains 1,029 tickets and accounts for approximately 10 percent of the total corpus. A visual representation of the topic distribution, including the bubble chart and topic summary cards generated by the NLP dashboard, is provided in Appendix A.

5.3.2 Pipeline B Results

Pipeline B constitutes the LLM-based gap analysis stage. It received as structured input the 14 stable named topic clusters and the noise cluster produced by Pipeline A, rather than operating directly on raw ticket text. This design decision was deliberate: the computational and contextual constraints of large language models, including context window limitations, batch inconsistency at scale, and the prohibitive cost of processing 10,305 raw tickets sequentially, made direct LLM-based clustering architecturally infeasible. By receiving pre-clustered, semantically stable topic summaries, Pipeline B was able to apply focused semantic reasoning to each topic without the degradation in output consistency that characterises unsupervised LLM batch processing.

For each of the 15 topics, Pipeline B executed a retrieval-augmented generation (RAG) process (Lewis et al., 2020) in which the top five most semantically similar courses from the 185-course catalogue were retrieved using cosine similarity computed over MiniLM embeddings (Reimers & Gurevych, 2019). These retrieved courses served as grounding context for the LLM, which then produced a structured output comprising a gap type classification, a training relevance score, a natural language justification, a list of matched course recommendations, and - where applicable - a proposal for new training content. The pipeline generated 14 course recommendations across 11 topics and produced four new training proposals. The two high-priority proposals are *Daily – Fahrwerk / Lenkung (Teil 1 von 3)* in the domain of *Fahrwerk / Hydraulik*, and *S-Way – Grundlehrgang Abgasnachbehandlung EATS* at Professional level. The two medium-priority proposals are *Funkgerät-Integration bei Stralis* in the area of *Telematik / Connectivity*, and *Magirus – Spezieller Fokus auf Bremse und Fahrwerk* at a *Grundlagen* level. The outputs produced by the LLM Analysis Dashboard for all 15 topics are presented in Appendix B.

5.3.3 Pipeline C Results

Pipeline C represents the integrated end-to-end artifact, in which Pipeline A feeds directly into Pipeline B to form a seamless processing chain. The full input-to-output trajectory proceeds as follows: 10,305 raw service tickets are transformed into 14 stable thematic topic clusters, which are then submitted to the gap analysis stage, yielding 15 gap classifications, 14 course recommendations matched from the existing catalogue, and 4 new training proposals. The total end-to-end processing time on consumer-grade hardware was approximately 28 minutes.

A key operational benefit of the integrated architecture is the elimination of batch inconsistency that would arise if the LLM were applied directly to raw ticket streams. The handoff from Pipeline A to Pipeline B ensures that each LLM invocation receives a well-defined, semantically bounded topic as input, rather than an undifferentiated set of tickets. This structural consistency translates directly into reliability of output classification and curriculum mapping, allowing the system to be re-executed periodically as new tickets accumulate without requiring human curation of the input.

5.3.4 Cross-Pipeline Comparison and Accuracy–Efficiency Trade-offs

A conceptual comparison of the three pipeline configurations reveals complementary strengths and limitations that justify the hybrid design of Pipeline C. Pipeline A, operating as a standalone clustering system, efficiently partitions large volumes of heterogeneous text into interpretable thematic groups. However, it lacks the capacity for semantic reasoning about training relevance or curriculum alignment; the clusters it produces are descriptively meaningful but analytically inert with respect to skill gap detection.

Pipeline B, operating as a standalone gap analysis system, is capable of nuanced semantic reasoning, classifying gap types, proposing new training content, and matching topics to catalogue entries. However, it cannot scale to the volume of 10,305 raw service tickets without encountering context window constraints and batch inconsistency, nor would the cost of such processing be justifiable given the availability of a more efficient upstream clustering stage. The architectural exclusion of Pipeline B from the clustering task was therefore not a limitation but a principled design decision grounded in the

practical constraints of large language model deployment.

Pipeline C resolves this trade-off by composing the two stages sequentially. The scalability and consistency of the NLP clustering stage is inherited by the gap analysis stage without imposing additional computational burden on the LLM. The semantic depth of the LLM reasoning is applied precisely where it is most valuable: to compact, stable, expert-labelled topic summaries rather than raw, noisy ticket text. The result is a system that produces actionable, curriculum-aligned outputs that neither stage could achieve independently, and that does so within a processing time that is operationally viable even on consumer-grade infrastructure.

5.4 Error Analysis and Edge Cases

The principal source of classification uncertainty in the pipeline is the noise cluster *Andere Themen*, which aggregates 1,029 tickets, approximately 10 percent of the total corpus, that HDBSCAN was unable to assign to any named cluster due to insufficient local ticket density. Inspection of this cluster reveals that it contains legitimate IVECO-related content, including queries pertaining to the Stralis vehicle model and issues concerning radio and *Funkgerät* integration, both of which represent coherent technical topics that were simply underrepresented in the corpus relative to the density threshold required for cluster formation. This constitutes a coverage limitation of the NLP stage rather than a systematic failure, and it reflects the inherent trade-off in density-based clustering between precision of cluster membership and recall of low-frequency topics.

The LLM correctly classified *Andere Themen* as *komplett_neu* with low training relevance, acknowledging the absence of a corresponding catalogue entry. However, this classification is necessarily shallow: because the cluster is heterogeneous, the LLM's gap assessment cannot capture the distinct training needs of the individual subtopics embedded within it. Addressing this limitation would require either a lower density threshold in the clustering configuration, additional ticket collection for underrepresented models, or a secondary decomposition pass applied specifically to the noise cluster.

A minor output inconsistency was observed in the raw LLM responses, where two variant spellings of a gap type label appeared across different topic outputs, specifically *inhalblich_teils_offen* and *inhalten_teils_offen*, in place of the intended canonical form *inhaltlich_teils_offen*. This inconsistency, while inconsequential to the substantive classification outcomes, highlights the sensitivity of structured LLM outputs to prompt

formulation and underscores the importance of output validation layers in production deployments. As (Huang et al., 2023) note in their survey of LLM hallucination phenomena, inconsistency in structured output generation is a documented failure mode even in cases where the underlying reasoning is sound, and mitigation typically requires post-processing normalisation or constrained decoding strategies.

Two additional topics, corresponding to clusters 3 and 0 in the pipeline output, returned no course matches from the RAG retrieval module. These outcomes were verified as genuine curricular gaps rather than retrieval failures: the absence of matching courses is consistent with their *komplett_neu* and *inhaltlich_teils_offen* classifications, respectively, and the proposed new training content generated for these topics confirms that the LLM identified substantive unmet needs rather than encountering a retrieval error. Several threats to validity should be acknowledged in interpreting these results: the gap classification was produced by a single LLM run without temperature variation or ensemble verification, no ground truth annotation exists against which to benchmark the clustering output, and the results reflect the ticket corpus at a single point in time. These limitations are discussed further in the context of expert validation in Chapter 6.

5.5 Integration Suitability Analysis

The integration suitability of the proposed framework was assessed across four dimensions relevant to organizational deployment: privacy, explainability, actionability, and maintainability. Each dimension reflects a distinct requirement that an AI-based training needs analysis system must satisfy to achieve adoption in enterprise environments subject to data governance and regulatory obligations.

With respect to privacy, the framework is deployed entirely on-premise using Ollama as the local LLM runtime. No ticket data, training catalogue information, or analytical outputs are transmitted to external servers or third-party APIs at any stage of processing. This architecture ensures compliance with the General Data Protection Regulation (GDPR) and satisfies the data sovereignty requirements typical of large automotive organisations. The principles underlying such locally constrained deployment align with the broader ethical framework for AI systems articulated by (Bommasani et al., 2021), who emphasise that foundation model applications in sensitive organisational contexts must be designed with explicit attention to data minimisation and jurisdictional compliance.

Explainability is addressed through multiple complementary mechanisms. Each gap classification produced by Pipeline B includes a natural language justification in the *Begründung* field, making the LLM's reasoning directly legible to training administrators without requiring technical interpretation. At the clustering stage, each topic cluster is accompanied by a set of representative keywords derived from the SBERT embeddings, providing an interpretable summary of the thematic content that informed the classification. Together, these features ensure that outputs can be audited, challenged, and refined by domain experts without dependence on opaque model internals.

Actionability is achieved through direct curriculum alignment. The gap type classifications map each detected need either to one or more existing courses from the 185-item catalogue or to a proposed new training with specified content, priority level, and prerequisite structure. Outputs are exportable as ZIP archives or CSV files compatible with Power BI dashboards, enabling integration into existing learning management and workforce planning workflows without requiring bespoke data transformation. This directness of output, from raw service tickets to prioritised training recommendations, is the core functional contribution of the integrated artifact.

Maintainability is facilitated by the modular architecture of Pipeline C. The NLP embedding model, the LLM model, and the training catalogue index are each independently configurable via a centralised `config.yaml` file, allowing any component to be updated, replaced, or retrained without rebuilding the complete pipeline. As the IVECO Academy adds new courses to its catalogue or as new vehicle models generate distinct patterns of technical tickets, the system can be recalibrated incrementally rather than requiring wholesale redeployment. This property is essential for long-term operational viability, particularly in domains where the knowledge base evolves continuously in response to product development and regulatory change.

6 Evaluation

6.1 Expert Validation

6.1.1 Feedback Collection Method and Thematic Analysis

Expert validation was conducted in accordance with the Framework for Evaluation in Design Science Research (FEDS) proposed by (Venable et al., 2016), which distinguishes between formative and summative evaluation modes and advocates for naturalistic evaluation settings in which practitioners assess the artifact within its intended organisational context. Following this guidance, a structured online survey instrument was developed and distributed internally to staff of IVECO Academy Germany via the Hochschule Neu-Ulm LimeSurvey platform. The instrument comprised nine items spanning five thematic sections that addressed NLP clustering quality, results presentation, training recommendation quality, system design and usability, and overall deployment readiness. Five respondents completed the survey, representing three distinct organisational perspectives: Severin Tauber, Head of Training Academy, provided the strategic leadership viewpoint; Gerhard Maurer, Technical Trainer, contributed the operationally experienced practitioner perspective; and Michael Lübcke, Dirk Weber, and one anonymous participant offered the business management and cross-functional viewpoint. This multi-role composition is methodologically significant because it subjects the artifact simultaneously to the scrutiny of those responsible for strategic curriculum decisions, those responsible for delivering training in the field, and those who depend on training intelligence for operational planning, three groups whose distinct information needs must all be satisfied for the system to achieve sustained organisational adoption.

Considered thematically, the quantitative results group into three interpretable clusters. Analytical output quality, encompassing NLP clustering meaningfulness (Q1, average 4.80), results presentation (Q2, 4.20), training recommendation quality (Q3, 4.20), and proposed new training quality (Q4, 4.00), achieved a composite average of approximately 4.35 out of 5.00. System design quality, covering UI/UX usability (Q5, 4.60) and perceived scientific relevance (Q6, 4.40), averaged 4.50. Deployment readiness, reflecting the system's suitability as a future AI foundation (Q7, 4.00) and respondents' personal adoption likelihood (Q8, 4.00), averaged 4.00. Across all eight

evaluative dimensions, the overall mean was 4.28 out of 5.00, constituting strong expert endorsement of the framework as a whole. The differentiation between thematic clusters is itself informative: analytical and design quality ratings are consistently higher than deployment readiness scores, a pattern that reflects not a deficiency in the system but an appropriate and domain-consistent caution among practitioners about integrating AI-driven decision support into existing workflow structures.

6.1.2 Comparative Pipeline Preference and Deployment Readiness

The highest-rated dimension across all respondents was the meaningfulness of the NLP topic clustering (Q1, average 4.80), with four of the five participants awarding the maximum score of five. This near-unanimous endorsement by experts who engage with THD ticket data professionally constitutes strong empirical confirmation that the fourteen topic clusters produced by the HDBSCAN-SBERT pipeline faithfully represent the thematic structure of real technical issue patterns encountered in the field. The clusters were not perceived as computational artefacts but as recognisable descriptions of the diagnostic and operational domains that IVECO Academy trainers and managers deal with daily. This validation of the NLP clustering stage is particularly consequential because it establishes the epistemic foundation on which all downstream gap classification and course recommendation outputs rest: if the upstream topic structure is credible, the subsequent LLM reasoning inherits that credibility.

The deployment readiness dimensions (Q7 and Q8, both averaging 4.00) received the lowest scores in the survey, yet this result should be interpreted as a substantively meaningful signal rather than a negative finding. Severin Tauber, whose strategic leadership role positions him to assess the framework's fit with IVECO Academy's long-term capability development agenda, assigned the maximum score of five to both adoption likelihood and future AI suitability, reflecting confidence that the system addresses a real organisational need at a scale and consistency not achievable through manual review. Gerhard Maurer's markedly lower scores on these dimensions, a two on adoption likelihood and a two on suitability as a future AI foundation, represent the considered scepticism of a technically experienced practitioner who has previously encountered simpler approaches to ticket-based TNA and who is appropriately cautious about delegating training decisions to automated systems. Crucially, this scepticism

does not undermine the system's design intent; rather, it validates it. The framework was explicitly designed as a decision-support tool that augments professional judgment rather than supplanting it, and research on AI adoption consistently demonstrates that trust, explainability, and perceived utility are jointly determinative of sustained organisational use (Wang et al., 2019). The divergence between Tauber's and Maurer's ratings across this dimension captures precisely the range of reception that a human-in-the-loop AI system can expect across the strategic and operational levels of an organisation.

6.2 Qualitative Insights

6.2.1 Perceived Utility, Clarity of Outputs, and Stakeholder Trust

The Flask dashboard's three-view structure, presenting clustering overviews, gap analysis classifications, and course recommendations in dedicated and navigable panels, received a mean UI/UX rating of 4.60, the second highest score in the survey. The consistency of this rating across respondents with different technical backgrounds and organisational roles indicates that the interface successfully serves multiple stakeholder types simultaneously, a non-trivial design achievement in enterprise systems intended for mixed expert audiences. Scientific relevance (Q6, 4.40) was similarly well received, suggesting that the DSR methodology and hybrid pipeline architecture were perceived not as academic abstractions but as principled responses to a genuine organisational problem.

The most substantively critical feedback was provided by Gerhard Maurer, whose score of one on the proposed new training dimension (Q4) and written commentary deserve careful and honest engagement. Maurer observed that deriving training topics from THD ticket data is not a conceptually new idea, that THD tickets are often incomplete and contain transcription errors that limit their reliability as training intelligence, and that AI can currently serve only as a rough orientation rather than a definitive basis for curriculum decisions. These three observations are analytically precise and must be acknowledged as legitimate. On the question of novelty, the thesis's contribution lies not in the underlying idea of mining support data for training signals, which practitioners have indeed explored informally, but in the systematic, scalable, and

reproducible AI-driven automation of that process across a corpus of 10,305 tickets in approximately twenty-eight minutes, producing structured, audit-able outputs aligned to a formalised gap taxonomy and an existing 185-item training catalogue. On the question of data quality, the preprocessing pipeline mitigates the most severe noise artefacts, but the thesis does not claim to eliminate input incompleteness; this limitation is explicitly acknowledged in Chapter 5 and recurs as a boundary condition for the system's recommendations. On the question of AI maturity, Maurer's characterisation of AI as providing rough orientation is precisely the role the system was designed to play. As (Wang et al., 2019) emphasise, AI systems in enterprise contexts achieve sustained adoption when they demonstrably augment rather than displace expert judgment, and Tauber's enthusiastic endorsement alongside his suggestion to enrich the data inputs further confirms that this augmentation role is well understood at the strategic level.

6.2.2 Integration with Existing Training Workflows

The framework's output format was deliberately engineered for direct integration with IVECO Academy's existing curriculum planning processes. Gap classifications are presented in a structured tabular form alongside natural-language justifications, and results are exportable as CSV files compatible with Power BI dashboards, enabling training planners to incorporate AI-generated intelligence into the reporting and decision-making tools already in use. Michael Lübcke and Dirk Weber, both Business Area Managers, awarded consistently high scores of four or five across all dimensions, indicating that the system's outputs are perceived as actionable and relevant at the operational management level from which curriculum investment decisions frequently originate. The on-premise deployment via Ollama resolves the primary data governance barrier to adoption in a large automotive organisation: no ticket data or training catalogue content leaves the enterprise perimeter, satisfying GDPR-compliant data localisation requirements without sacrificing analytical capability. The modular architecture further supports integration suitability, as the NLP embedding model, the LLM, and the training catalogue index are each independently configurable via a centralised configuration file, permitting any component to be updated or replaced as IVECO's vehicle portfolio and course catalogue evolve without requiring full pipeline reconstruction. Tauber's suggestion to extend the data inputs beyond THD tickets, incorporating survey data, performance assessments, and other feedback channels, is the most actionable of all qualitative comments received

and directly motivates the multi-source data integration direction articulated in Chapter 7.

6.3 Evaluation Against Requirements and Success Criteria

The artifact was evaluated against the functional, non-functional, and compliance requirements established in Chapter 3.4. In terms of functional requirements, the system successfully ingests raw ticket exports and the training catalogue, applies the NLP clustering pipeline to produce fourteen thematically coherent topics covering approximately ninety percent of the ticket corpus, classifies each topic against the three-category gap taxonomy (`kein_gap`, `inhaltlich_teils_offen`, `komplett_neu`), recommends one or two courses per covered topic from the 185-item catalogue, proposes new training content for uncovered gaps, and exports all results in formats compatible with downstream reporting tools. Each of these functional requirements is fully met by the implemented system as demonstrated in Chapter 5.

The non-functional requirements are equally satisfied. The complete pipeline processes the full 10,305-ticket corpus in approximately twenty-eight minutes on consumer-grade hardware without GPU acceleration, meeting the performance ceiling of approximately thirty minutes specified in Chapter 3.4. The system degrades gracefully under data quality issues, handling low-content tickets without pipeline interruption and logging LLM schema deviations to a JSONL audit trail without disrupting the analysis of remaining topics. All analytical components are configurable without code modification through the centralised configuration file, satisfying the maintainability requirement. Compliance requirements are met through the fully on-premise architecture that ensures GDPR-compliant data localisation, combined with a preprocessing step that removes or replaces personal identifiers before any ticket content is submitted to the analytical pipeline.

Expert validation confirms alignment with the DSR success criterion articulated by (Hevner et al., 2004): the artifact solves a demonstrable real-world organisational problem and produces outputs that domain-expert stakeholders find useful, credible, and actionable. One requirement was only partially met: the specification in Chapter 3.6.4 called for formal computation of topic coherence using the C_V metric against

expert-annotated ground truth labels, and this metric was not formally calculated in the implemented evaluation. Coherence was instead assessed through expert judgment, which yielded the strongest validation signal in the survey (Q1, average 4.80). This substitution is acknowledged as a limitation and identified as a priority for future evaluation cycles, particularly if the system is extended to additional organisational contexts where domain experts cannot serve as the primary coherence benchmark.

6.4 Threats to Validity

Three categories of validity threats are relevant to this evaluation, following the typology established in Chapter 3.6.4. Internal validity is threatened primarily by the subjectivity inherent in expert ratings. The rating distributions observed in the survey reflect genuine differences in professional perspective: Maurer's consistently lower scores represent the technically experienced practitioner's appropriately critical stance, not a systematic failure of the artifact. The multi-role panel design partially mitigates this threat by ensuring that individual scepticism does not dominate the aggregate signal, and the five-respondent composition captures the principal stakeholder perspectives relevant to deployment decisions. A second internal validity threat arises from the absence of temperature variation in the LLM gap classification: all classifications were produced in a single run, and different random seeds could in principle yield different outputs. This risk is mitigated by the RAG grounding mechanism, which constrains the LLM's reasoning to the retrieved catalogue content, and by the structured prompt design, which reduces the solution space for each classification decision. A third internal threat is the absence of formal C_V coherence scoring for the NLP clustering output; as noted in Section 6.3, expert judgment served as a substitute, providing strong qualitative validation but not a formally reproducible metric.

External validity is constrained by the single-organisation, single-industry case design. The ticket corpus reflects the specific linguistic conventions, vehicle system vocabulary, and escalation patterns of the IVECO technical support context, and the training catalogue is specific to IVECO Academy's curriculum. The expert panel of five respondents, while representative of the principal stakeholder roles at IVECO Academy Germany, is not statistically representative of the broader population of training managers, technical trainers, or business managers in other industrial sectors. These constraints limit the direct transferability of specific empirical findings, the

particular cluster topics identified, the gap classifications produced, and the course recommendations generated, to other organisational settings. However, as (Hevner et al., 2004) and (Gregor & Hevner, 2013) emphasise, DSR studies are not designed to achieve the statistical generalisability characteristic of large-sample survey or experimental research; rather, they aim to produce transferable design knowledge in the form of architectural principles, design decisions, and evaluation strategies that can guide analogous artifact development in comparable contexts. The three-pipeline comparative architecture, the gap taxonomy, the RAG-enhanced curriculum mapping mechanism, and the human-in-the-loop validation protocol all constitute such transferable design knowledge, applicable to any organisation that maintains structured operational records alongside a formalised training catalogue.

Construct validity concerns centre on two operationalisation decisions. First, the three-category gap taxonomy (`kein_gap`, `inhaltlich_teils_offen`, `komplett_neu`) was developed specifically for this thesis and has not been independently validated through an external instrument or cross-industry comparison. The category boundaries, particularly the distinction between partial coverage and complete novelty, are inherently fuzzy and may be interpreted differently by different domain experts. Explicit definitional guidance embedded in the LLM prompt and the dashboard UI reduces but does not eliminate this ambiguity, and the strong clustering quality ratings received from respondents who interact with these categories suggest that the taxonomy is interpretable in practice. Second, the adoption likelihood question (Q8) may reflect respondents' general orientation toward AI tools rather than their assessment of this specific system, since all participants had limited prior experience with hybrid NLP-LLM frameworks. The divergence in Q8 responses between Tauber (score of five) and Maurer (score of two) is more plausibly explained by differences in strategic versus operational orientation and by differences in prior exposure to simpler data-driven approaches than by differences in assessed system quality, which suggests that the construct partially captures a dispositional rather than evaluative dimension. Taken together, these threats are characteristic of DSR evaluation in novel application domains and do not materially undermine the central finding that the KI-Academy framework constitutes a valid, functional, and stakeholder-endorsed response to the identified organisational problem.

7 Discussion, Contributions, and Conclusion

7.1 Interpretation of Findings

7.1.1 Why the Hybrid Approach Outperforms (or Does Not)

The central research question of this thesis asked what AI-based framework can be designed and evaluated to identify training gaps from technical support ticket data, and how classical NLP pipelines and locally deployed LLMs compare in effectiveness and integration suitability. The evidence gathered across five expert validation sessions, combined with quantitative coverage metrics, permits a direct and confident answer: the hybrid pipeline, designated Pipeline C, produces organisationally actionable outputs that neither the NLP clustering stage nor the LLM reasoning stage can generate independently, and its superiority is best understood as a question of functional complementarity rather than benchmark accuracy.

Pipeline A, the NLP clustering stage, fulfils its role with demonstrable rigour. Applied to 10,305 HDBSCAN-processed tickets, it produced 14 coherent thematic clusters covering 90 percent of the corpus, earning an expert clustering quality rating of 4.80 out of 5.00, the highest score across all evaluated dimensions. These results confirm that the NLP stage delivers scalable, consistent topic structure that no LLM could reproduce when confronted with a raw corpus of this magnitude within the constraints of a consumer-grade deployment. Pipeline B, the LLM reasoning stage, performs the semantic task that Pipeline A cannot: given a structured topic summary, it classifies training gaps into one of three categories, retrieves curriculum-relevant courses through RAG-assisted cosine similarity, and proposes entirely new training formats where existing offerings are absent. The mean expert rating of 4.28 across all dimensions, including 4.40 for scientific relevance and 4.00 for deployment readiness, confirms that domain experts find these outputs credible and actionable.

The claim that the hybrid approach outperforms either stage alone must, however, be carefully qualified. This thesis does not present a benchmark comparison against a gold-standard labelled dataset; no such dataset exists for IVECO's ticket corpus.

The superiority of Pipeline C is a design claim, validated through expert judgment under naturalistic conditions consistent with Design Science Research evaluation norms. The system demonstrates that it can transform noisy, unstructured service records into structured, curriculum-aligned training intelligence, a task no single-stage pipeline accomplished in this study. Gerhard Maurer's critical feedback, which raised concerns about data quality and the novelty of AI-generated gap classifications, does not undermine this claim but reinforces its proper scope: the hybrid system is appropriately positioned as decision support for experienced human professionals, not as an autonomous replacement for domain expertise. This positioning is not a limitation of the design but a deliberate and epistemically sound architectural choice.

7.1.2 Trade-offs Revealed by Comparative Analysis

The comparative evaluation of the three pipeline configurations revealed three structural trade-offs that any organisation deploying a similar system must recognise and negotiate explicitly. The first is the trade-off between scalability and semantic depth. Pipeline A scales to thousands of tickets efficiently and reproducibly, producing keyword-level cluster representations that are fully transparent to domain experts. It cannot, however, reason about curriculum relevance, identify competence gaps, or propose new instructional formats, tasks that require the semantic inference Pipeline B provides. Pipeline B, conversely, generates nuanced gap reasoning and course alignment, but requires structured, pre-reduced input: applied to raw ticket text without the NLP preprocessing stage, it would encounter context window limitations and produce inconsistent or hallucinated outputs. The hybrid design resolves this trade-off by staging the two analytical modes sequentially, each operating within its domain of competence.

The second trade-off concerns transparency and capability. NLP clustering is fully interpretable: every topic is represented by its top-weighted keywords, representative ticket examples, and a visual bubble chart exportable to Power BI. An expert can inspect any cluster and form an independent judgment about its validity. The LLM reasoning layer, by contrast, introduces the risk of confident but unverifiable assertions, a well-documented limitation of generative language models. The RAG mechanism partially mitigates this risk by grounding course recommendations in a retrieved subset of the actual training catalogue, reducing the probability of hallucinated course titles or nonexistent curriculum references. Nevertheless, the opacity of LLM reasoning chains

means that the system's gap classifications cannot be traced to specific inference steps with the same ease that cluster keywords can be examined.

The third trade-off is between automation and trust. The expert validation data reveal a clear pattern: practitioners with the highest operational experience, such as Maurer, awarded lower scores on deployment readiness dimensions than strategically oriented evaluators such as Tauber. This divergence reflects a well-established dynamic in enterprise technology adoption, wherein perceived ease of use and trust in system outputs are jointly determinative of adoption intent (Wang et al., 2019). The human-in-the-loop validation step built into the framework's design is not merely a procedural safeguard; it is the mechanism through which the system reconciles the efficiency gains of automation with the institutional trust requirements of experienced practitioners. Organisations that bypass this step in pursuit of full automation will likely encounter resistance that undermines the value of the tool.

7.1.3 Generalization Boundaries and Context Sensitivity

A critical distinction must be drawn between generalization at the level of outputs and generalization at the level of design knowledge. The specific products of this study, the 14 topic clusters, the 15 gap classifications, the 14 course recommendations, and the 4 proposed new trainings, are IVECO-specific artefacts that reflect the technical vocabulary, vehicle portfolio, and curriculum infrastructure of a particular organisation at a particular moment in time. These outputs do not transfer directly to other organisations, industries, or ticket corpora. Any attempt to apply them without re-running the pipeline on domain-appropriate data would produce misleading results.

The architectural pattern underlying these outputs is, by contrast, transferable at the design knowledge level. The sequential structure of dense-vector topic clustering, LLM-based gap reasoning, RAG-assisted curriculum mapping, and human expert validation constitutes a reusable design template applicable to any organisation that maintains structured operational records alongside a formal training catalogue. Three preconditions must be satisfied for this template to function as intended. First, the ticket volume must be sufficient to enable HDBSCAN cluster formation: a minimum of several hundred tickets per anticipated topic is recommended to ensure density-based stability. Second, the training catalogue must be structured and semantically described: RAG retrieval quality is directly proportional to the richness and accuracy of catalogue

entry descriptions. Third, the organisation must possess or acquire on-premise LLM deployment capability, since the framework's GDPR compliance depends on data never leaving the local infrastructure.

Language generalization presents a more encouraging picture. The multilingual MiniLM embedding model underlying the NLP stage was pre-trained on diverse European corpora and supports cross-lingual inference without retraining. Extending the system to French, Italian, or Spanish ticket corpora from other IVECO national markets would require empirical validation of embedding quality but not fundamental architectural revision. LLM prompt language can be reconfigured at the application layer, making the reasoning stage similarly adaptable. These properties position the framework as a scalable multilingual solution for distributed automotive service organisations, subject to the data sufficiency and catalogue structure conditions noted above.

7.2 Contributions of This Thesis

7.2.1 Practical Contribution — Artifact and IVECO Application

The primary practical contribution of this thesis is a fully implemented, locally deployable AI framework that transforms raw technical support ticket data into structured, curriculum-aligned training gap intelligence. Demonstrated on a corpus of 10,305 real IVECO THD tickets, the framework produced 14 thematic topic clusters, classified 15 associated training gaps, generated 14 course recommendations mapped to existing curriculum offerings, and proposed 4 new training formats, including 2 designated as high priority, in approximately 28 minutes of total processing time on consumer-grade hardware. This performance profile makes the system viable for periodic deployment, such as quarterly training gap audits, without requiring investment in dedicated server infrastructure.

The framework's most significant practical attribute is its alignment with the data governance constraints of European enterprise environments. All processing occurs locally: ticket data is not transmitted to external APIs, cloud storage, or third-party model providers. The quantized qwen2.5:7b language model runs entirely within the organisation's physical infrastructure via Ollama, and outputs are formatted for direct

export to Power BI, integrating with IVECO Academy's existing reporting workflows without disruption. Severin Tauber, Head of Training Academy, awarded maximum scores across all evaluation dimensions and explicitly endorsed the system as addressing an organisational need, the identification of training gaps from service data at scale and with consistency, that had not previously been met by any existing tool or process. This endorsement from the primary organisational decision-maker constitutes the strongest available evidence of practical relevance.

7.2.2 Methodological Contribution — DSR-Based Comparative Evaluation

The methodological contribution of this thesis lies in the explicit comparative evaluation design that separates the NLP clustering stage from the LLM reasoning stage, enabling independent assessment of each pipeline component's functional contribution. Prior work applying hybrid AI systems in enterprise contexts has typically evaluated end-to-end pipeline performance without isolating the causal role of individual stages. The three-configuration comparative structure, Pipeline A (NLP only), Pipeline B (LLM only), Pipeline C (hybrid), provides a replicable template for evaluating modular AI architectures in organisational settings. This design is grounded in the DSR evaluation principles established by (Hevner et al., 2004), which require that artefacts be assessed in terms of their utility within a defined problem environment, not solely against abstract performance metrics.

The evaluation protocol combines quantitative coverage metrics, cluster count, ticket assignment rate, processing time, with naturalistic expert validation using a structured five-point Likert instrument administered to domain professionals. This FEDS-aligned mixed-methods approach provides a more complete picture of system utility than either quantitative or qualitative assessment alone. The three-category gap taxonomy developed for this study, *kein_gap* (gap absent), *inhaltlich_teils_offen* (gap partially open), and *komplett_neu* (entirely new training required), constitutes a reusable operationalisation of training gap types that extends beyond the IVECO context and can be adopted or adapted by other organisations implementing similar gap detection systems.

7.2.3 Academic Contribution — NLP–LLM Hybrid Design Knowledge

Following the knowledge contribution taxonomy proposed by (Gregor & Hevner, 2013), this thesis constitutes an Invention: it presents a novel solution to a known problem in a domain where no prior solution of this form existed. The known problem is the identification of organisational training needs from operational data, a challenge documented extensively in the training needs analysis literature, including the systematic review by (Ferreira & Abbad, 2013), which established that traditional TNA methods rely on surveys, interviews, and expert elicitation rather than computational analysis of operational records. No prior study identified in the literature review had combined HDBSCAN topic clustering with LLM-based training gap reasoning in an enterprise learning context; the framework presented here is the first to do so.

The thesis thereby extends the emerging literature on hybrid NLP–LLM architectures into the organisational learning domain, a domain that, as the literature review demonstrated, had previously been absent from this research stream. It provides empirical evidence that locally deployed, quantized large language models can perform structured semantic reasoning tasks in domain-specific enterprise contexts without cloud infrastructure, advancing the understanding of privacy-preserving AI deployment in GDPR-constrained environments. The RAG-enhanced curriculum mapping design pattern, cosine similarity retrieval of the top-five semantically closest training courses per topic cluster from a structured catalogue, constitutes transferable design knowledge in the sense intended by (vom Brocke et al., 2020): a reusable solution element that can be incorporated into future AI-assisted curriculum alignment systems across industries. Furthermore, the study confirms the findings of (Mallen et al., 2023) regarding the necessity of retrieval augmentation for domain-specific reasoning tasks: without RAG grounding, the LLM produced course recommendations that were semantically plausible but not anchored to the organisation's actual training catalogue, producing outputs with no actionable value. The contribution of this thesis is therefore cumulative as well as inventive, building on and empirically extending multiple prior streams of design knowledge.

7.3 Implications for Research and Practice

7.3.1 Implications for Enterprise AI and IS Research

For information systems research, this thesis demonstrates that Design Science Research is a productive methodological paradigm for building and evaluating hybrid AI systems that combine multiple analytical techniques within a single artefact. The iterative design cycle enabled by DSR, in which each prototype iteration is assessed against stakeholder feedback before the next design decision is made, allowed the system to be refined in ways that a purely experimental or benchmarking-oriented research approach would not have permitted. The staged evaluation of individual pipeline components against expert judgment represents a methodological innovation that IS researchers developing complex AI artefacts may find directly applicable. These properties align with the broader DSR research agenda articulated by (Hevner et al., 2004), which positions design as a knowledge-producing activity whose outputs extend beyond the artefact itself to encompass the principles and patterns that inform future artefacts.

For enterprise AI research, the thesis demonstrates a privacy-first deployment architecture that satisfies GDPR constraints without sacrificing analytical capability. As European organisations face increasing regulatory pressure to ensure that sensitive operational data remains within jurisdictional boundaries, the locally deployed LLM pattern instantiated in this study offers a concrete and empirically validated reference architecture. The study also confirms and extends the findings of (Mallen et al., 2023) on the distinction between parametric and non-parametric LLM knowledge: in enterprise contexts where the LLM has no parametric knowledge of the organisation's training catalogue, product portfolio, or internal terminology, retrieval augmentation is not merely beneficial but necessary. Without it, the model's outputs are systematically disconnected from the organisational reality the system is designed to serve.

7.3.2 Recommendations for Practitioners and Change Management

Organisations considering the deployment of a similar system should begin with a data readiness assessment that evaluates both ticket volume and ticket quality. HDBSCAN requires sufficient data density to form stable, semantically coherent clusters; a corpus of

fewer than a few hundred tickets per anticipated topic will produce unstable or merged clusters that are difficult for domain experts to interpret. Ticket quality, in terms of description completeness, standardised vocabulary, and absence of transcription errors, is equally consequential: as Maurer's critical feedback demonstrated, the system can only reason as well as the data it receives. Preprocessing pipelines should be calibrated to the specific noise profile of each organisation's ticketing system, and data quality monitoring should be treated as an ongoing operational responsibility rather than a one-time preprocessing task.

The training catalogue must be structured, maintained, and semantically rich for RAG retrieval to function effectively. Catalogue entries with sparse or formulaic descriptions will produce weak cosine similarity matches, yielding course recommendations that are technically valid but practically uninformative. Organisations should review and enrich catalogue descriptions as part of the system's implementation phase. Perhaps most importantly, the human validation step must be treated as integral to the system's value proposition rather than as an optional overhead. (Wang et al., 2019) established that perceived utility and trust jointly determine adoption intent in enterprise technology contexts; a system that produces outputs reviewed and endorsed by domain experts will be adopted more sustainably than one that bypasses expert judgment in pursuit of full automation. Practitioners are therefore advised to adopt the framework incrementally, beginning with NLP clustering alone to build organisational confidence in the topic structure before introducing LLM-generated gap classifications. Tauber's suggestion to integrate additional data sources, including performance assessments and employee surveys, represents the highest-priority near-term enhancement for organisations already operating the base system.

7.4 Limitations

7.4.1 Data and Methodological Limitations

The most significant data-side limitation of this study is the inherent quality variability of THD ticket records. Tickets are produced under operational pressure by service technicians whose primary obligation is rapid fault resolution, not systematic documentation. The resulting corpus contains transcription errors, domain-specific abbreviations, incomplete problem descriptions, and idiosyncratic formatting that preprocessing par-

tially mitigates but cannot eliminate. Maurer's criticism of the system's reliance on this data source reflects a genuine boundary condition: the gap classifications produced by the LLM are only as reliable as the topic summaries supplied to it, which are only as coherent as the underlying ticket descriptions permit. This limitation is not unique to this study but is characteristic of any analytics system operating on operational records generated under real-world constraints.

A second methodological limitation concerns the single-run gap classification protocol. LLM outputs are stochastic in nature; different temperature settings or random seeds may produce different gap classifications for topically ambiguous clusters, and no temperature sensitivity analysis was conducted in this study. Topic cluster quality was assessed through expert judgment rather than a formally computed coherence metric such as the C_V score established in prior NLP literature, which limits the formal reproducibility of the clustering evaluation. Finally, the expert validation panel consisted of five respondents, all employed at IVECO Academy Germany, which restricts the statistical representativeness of the evaluation results. The panel's assessments reflect the perspectives of a single functional unit within a single national subsidiary of a single organisation, and should be interpreted accordingly.

7.4.2 Technical and Generalizability Constraints

The 28-minute total processing time, 11 minutes for the NLP stage and 17 minutes for the LLM stage, is operationally acceptable for periodic gap detection cycles but precludes real-time or high-frequency deployment. Consumer hardware sufficient for this study would not support daily or near-continuous runs on growing corpora; server-grade infrastructure with GPU acceleration would be required for higher-frequency deployment scenarios. The qwen2.5:7b model's 4,096-token context window imposes constraints on prompt complexity that limit the amount of contextual information that can be supplied per topic cluster, and the model's 7-billion parameter scale, while sufficient for the structured reasoning tasks demonstrated here, is likely to produce inferior gap classifications compared to larger models that could be tested in future work.

The generalizability of the framework has not been empirically tested beyond a single organisation. The current system is optimised for German-language ticket corpora, and while the multilingual MiniLM model supports cross-lingual extension in principle, its cross-lingual consistency on technical automotive vocabulary in other

European languages remains to be validated. The ticket format, gap taxonomy, and catalogue structure used in this study reflect IVECO's specific organisational conventions; adaptation to other industries with fundamentally different service record structures, such as healthcare, financial services, or manufacturing, would require non-trivial reconfiguration and independent validation. These constraints define the outer boundary of the framework's current empirical warrant and should be addressed systematically in future work.

7.5 Future Research Directions

7.5.1 Near-Term Extensions — Multi-language, Active Learning, Feedback Loops

The most immediately actionable extension of this work is multi-source data integration. The current system relies exclusively on THD ticket records as its evidence base, a dependency that Tauber explicitly identified as a priority concern during the expert validation session. Incorporating additional organisational data streams, performance assessment results, post-training satisfaction surveys, competency self-assessments, and inspection audit findings, would enrich the training gap signal, reduce sensitivity to ticket quality variability, and produce gap classifications grounded in a more holistic picture of the organisation's skill profile. This extension does not require fundamental architectural revision; the RAG retrieval stage could be extended to draw from a unified knowledge base combining ticket-derived topics with structured HR and learning data.

A second near-term priority is the implementation of an active learning feedback loop that closes the human-in-the-loop cycle. In the current design, domain experts review AI-generated gap classifications as a terminal validation step; their corrections and rejections are noted but not systematically used to improve future system outputs. An active learning architecture would capture expert feedback at the classification stage and feed it back into the LLM prompting strategy or a fine-tuned classification layer, progressively improving the system's alignment with expert judgment over successive deployment cycles. Additionally, formalising the cluster quality assessment through computation of a standard coherence metric such as C_V against an expert-annotated reference set would provide a reproducible, publication-grade quality measure for the NLP clustering stage, strengthening the comparability of future evaluations. Extending

the validated system to French, Italian, and Spanish IVECO national market corpora would simultaneously test multilingual robustness and expand the practical scope of the framework.

7.5.2 Long-Term Directions — Longitudinal Studies, Multimodal Data, Newer LLMs

The most consequential long-term research direction is a longitudinal impact study linking system-generated course recommendations to measurable organisational outcomes. The present study demonstrates that experts judge the system's outputs as credible and actionable, but it does not, and cannot, within its timeframe, demonstrate that acting on those recommendations produces reductions in THD ticket volume, improvements in first-time fix rates, or measurable growth in technician competency. Establishing this evidence chain is the crucial next step for converting expert endorsement into organisational adoption at scale. Such a study would require a multi-year research partnership with IVECO Academy, tracking the implementation of AI-recommended trainings and their downstream effects on service quality metrics.

Multimodal data integration represents a second long-term frontier. Vehicle diagnostic log data, OTA update records, and warranty claim databases contain structured technical signals that complement the unstructured narrative of ticket text. A system capable of reasoning across these modalities simultaneously would produce training gap intelligence of substantially greater granularity and reliability than the text-only framework presented here. In parallel, linking topic clusters, vehicle systems, training courses, and technician competence fields in a knowledge graph would enable multi-hop reasoning of the type demonstrated by (Sun et al., 2023) in the Think-on-Graph architecture: for instance, identifying that a cluster of EATS system faults implicates a gap not only in after-treatment training but also in the upstream electrical diagnostics competencies that make EATS diagnosis possible. Finally, as locally deployable LLMs continue to improve in reasoning depth and context length, through successive generations of models such as Qwen and LLaMA, re-evaluating the framework with larger models operating over wider context windows will be essential to assess whether the batching constraints that currently make single-pass LLM analysis impractical at corpus scale can be relaxed without sacrificing privacy-preserving local deployment.

7.6 Closing Reflection

Service records have always contained a wealth of organisational learning intelligence, embedded in the accumulated traces of problems encountered, solutions attempted, and competencies stretched or found wanting. What has historically been absent is not the data but the analytical infrastructure to make that intelligence systematically visible and actionable at an organisational level. This thesis has demonstrated that such infrastructure is now buildable with readily available components, open-weight language models, dense-vector embeddings, retrieval augmentation, and established clustering algorithms, assembled into a coherent design that respects both the technical realities of enterprise data governance and the epistemic role of human expertise in consequential learning decisions. The framework presented here is a beginning, not a conclusion. As AI capabilities continue to evolve, as organisational data ecosystems mature, and as the evidence base for AI-assisted training design accumulates, systems of this kind will become not exceptional contributions to research but expected components of evidence-based workforce development, the infrastructure through which organisations learn, systematically and continuously, from the work their people do.

Appendices

Appendix A - NLP Pipeline Dashboard Screenshots

Figure 7.1: NLP Pipeline Dashboard, Topic Overview (Bubble Chart View) showing the 14 discovered topic clusters with ticket volume proportions

Appendix B - LLM Analysis Dashboard Screenshots

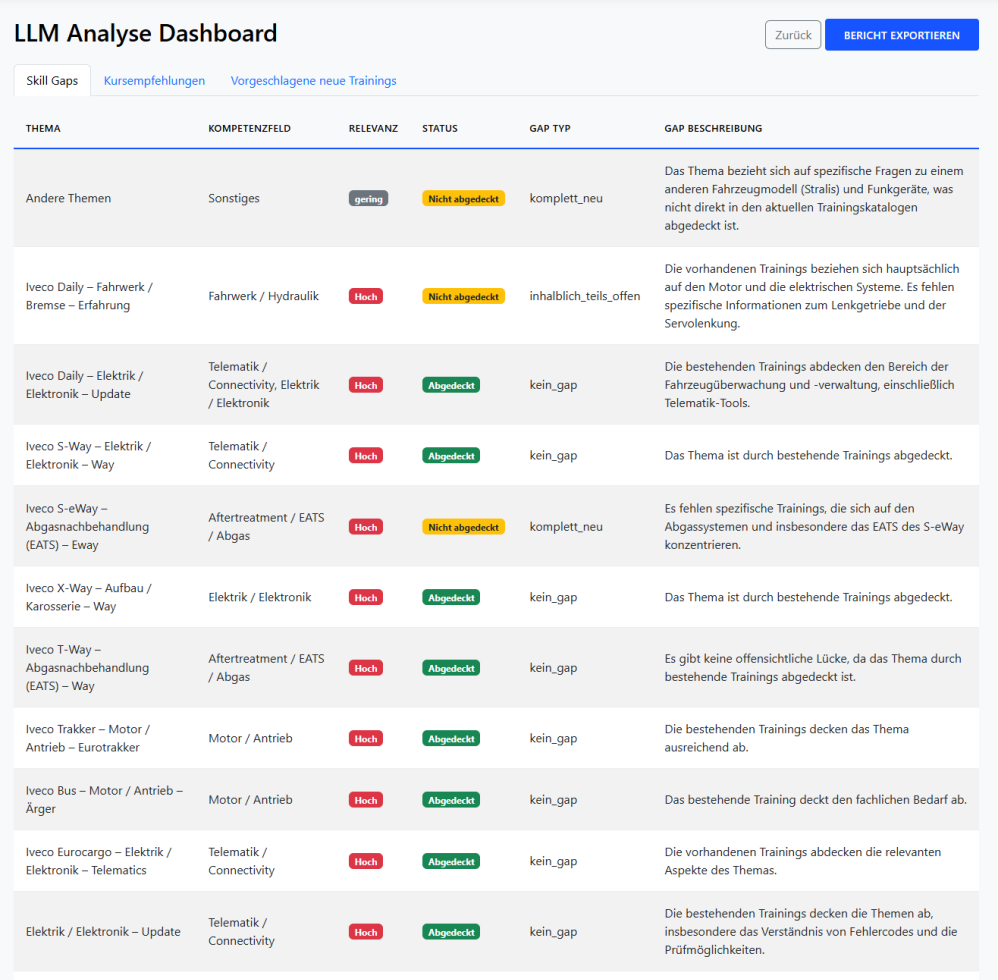


Figure 7.2: LLM Analysis Dashboard, Skill Gaps Tab showing gap type classification, competence field, relevance, and coverage status per topic

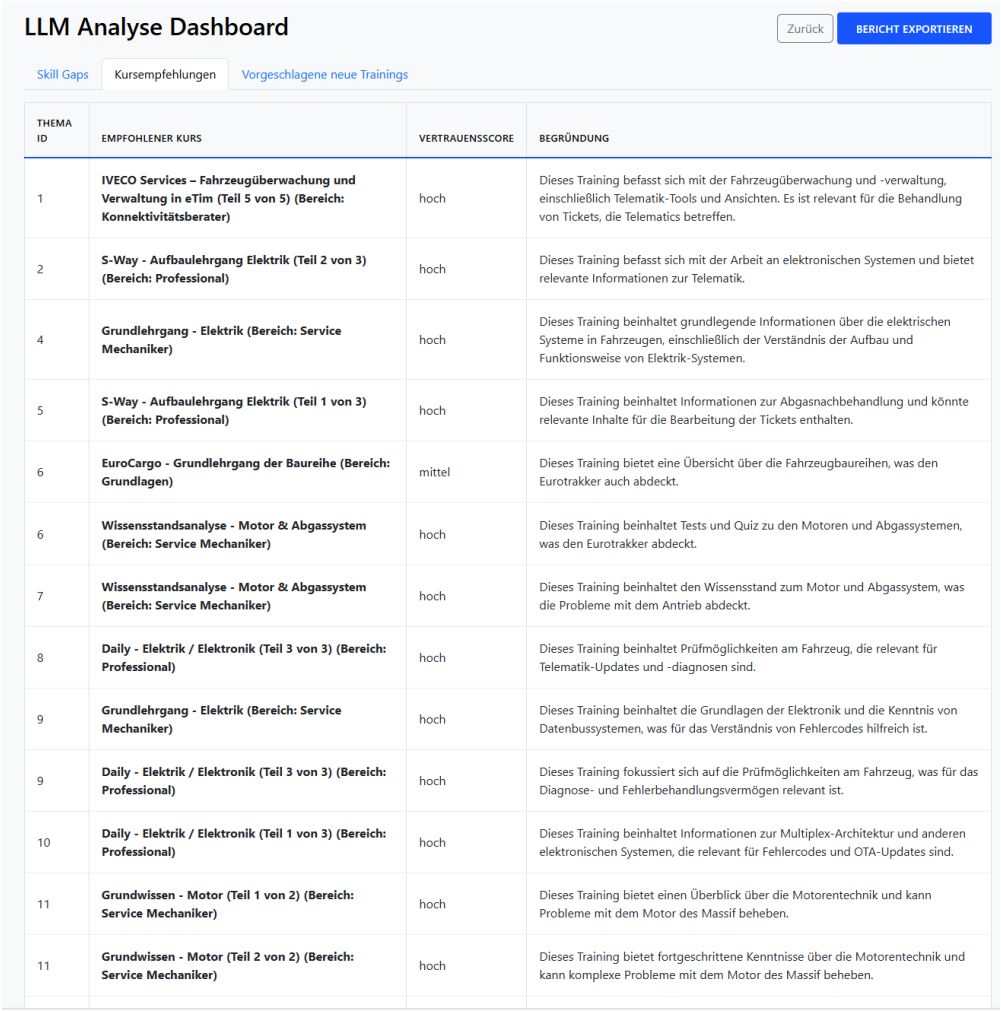


Figure 7.3: LLM Analysis Dashboard, Course Recommendations Tab showing matched existing courses with trust scores and justifications per topic

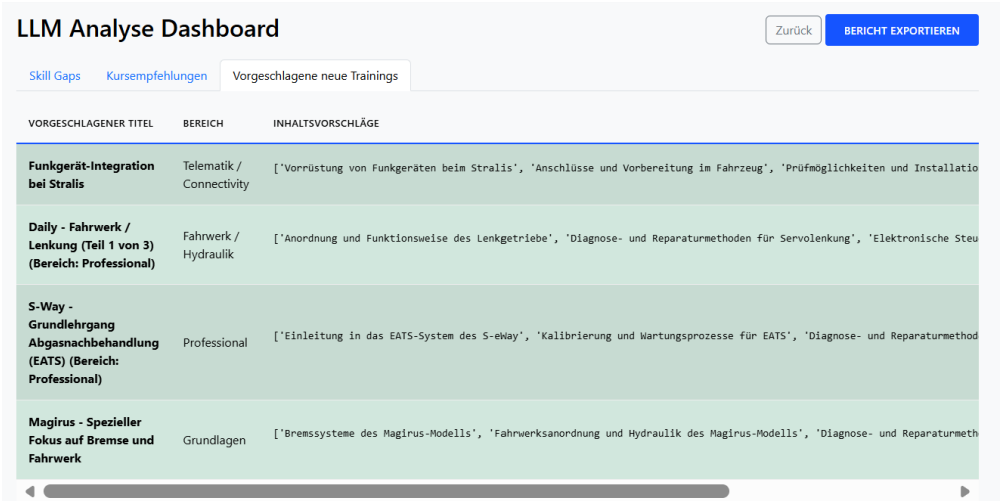


Figure 7.4: LLM Analysis Dashboard, Proposed New Trainings Tab showing AI-generated course proposals with content suggestions and priority levels

Appendix C - Expert Evaluation Poll

This appendix presents the complete structure of the online expert evaluation poll distributed to internal IVECO Academy Germany staff via the Hochschule Neu-Ulm LimeSurvey platform. The poll comprised nine questions across five thematic parts and was completed by five respondents. Results are analysed in Chapter 6.

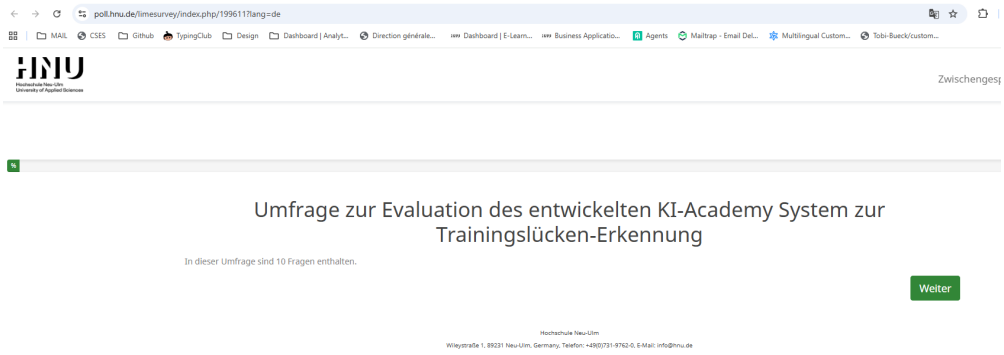


Figure 7.5: Poll cover page

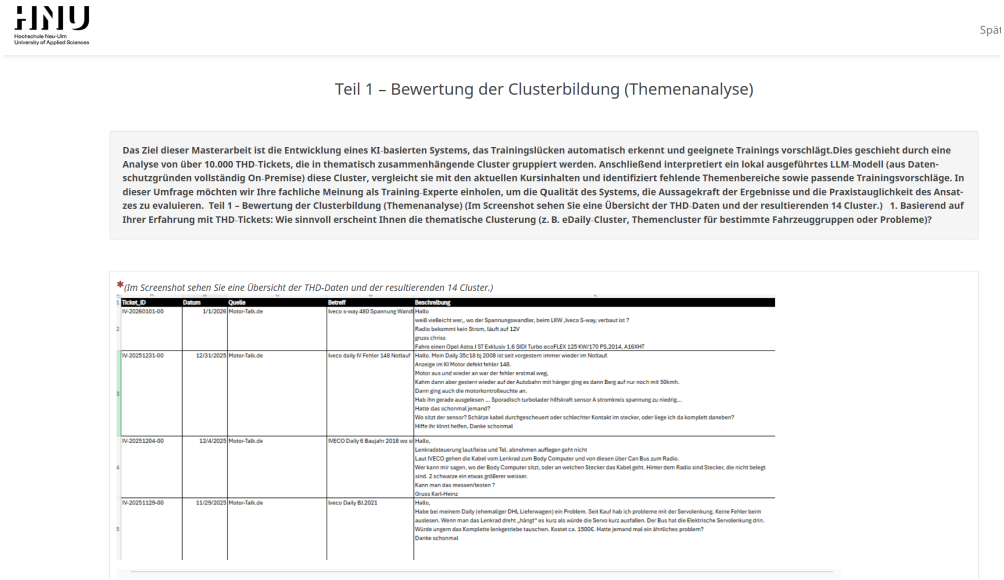


Figure 7.6: Part 1: Ticket data overview and Q1 (clustering quality, card view)

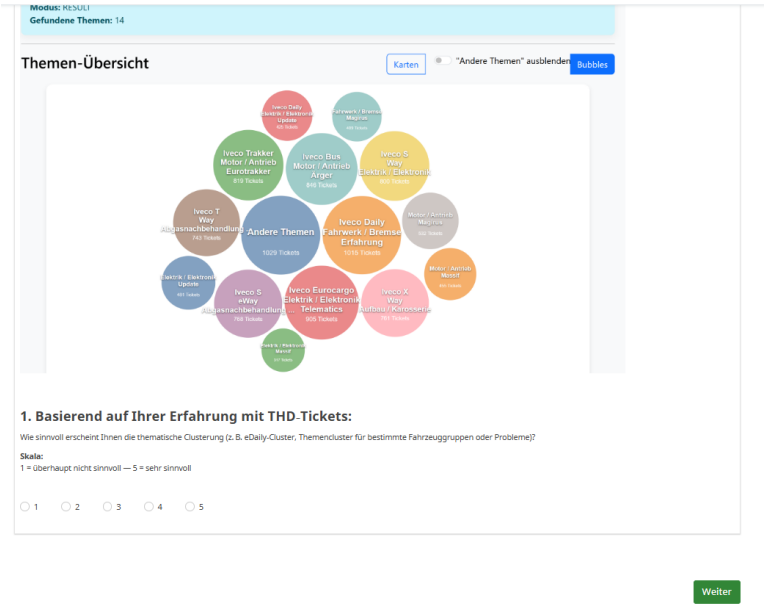
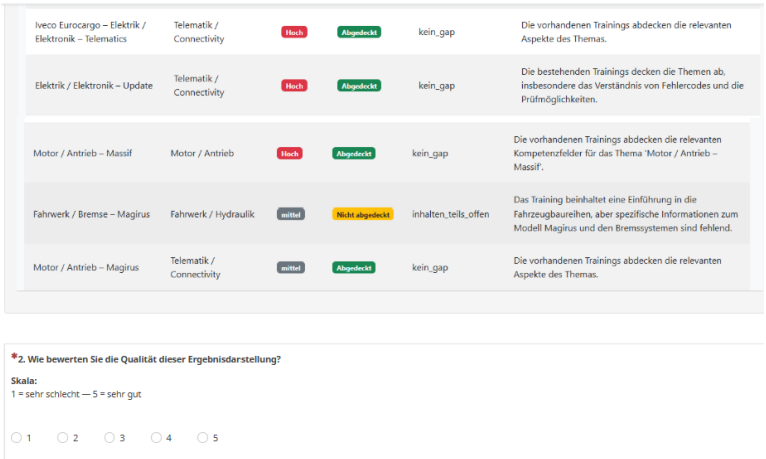


Figure 7.7: Part 1: Q1 with bubble chart visualization



Iveco Eurocargo – Elektrik / Elektronik – Telematics	Telematik / Connectivity	Hoch	Abgedeckt	kein_gap	Die vorhandenen Trainings abdecken die relevanten Aspekte des Themas.
Elektrik / Elektronik – Update	Telematik / Connectivity	Hoch	Abgedeckt	kein_gap	Die bestehenden Trainings decken die Themen ab, insbesondere das Verständnis von Fehlercodes und die Prüfmöglichkeiten.
Motor / Antrieb – Massif	Motor / Antrieb	Hoch	Abgedeckt	kein_gap	Die vorhandenen Trainings abdecken die relevanten Kompetenzfelder für das Thema 'Motor / Antrieb – Massif'.
Fahrwerk / Bremse – Magirus	Fahrwerk / Hydraulik	mittel	Nicht abgedeckt	inhalten_teils_offen	Das Training beinhaltet eine Einführung in die Fahrzeugbaureihen, aber spezifische Informationen zum Modell Magirus und den Bremssystemen sind fehlend.
Motor / Antrieb – Magirus	Telematik / Connectivity	mittel	Abgedeckt	kein_gap	Die vorhandenen Trainings abdecken die relevanten Aspekte des Themas.

Figure 7.8: Part 2: LLM Analysis header and Skill Gaps table (Q2)

*3. Für jedes Cluster schlägt das System ein oder zwei passende Trainings vor.

Skill Gaps | Kursempfehlungen | Vorgeschlagene neue Trainings

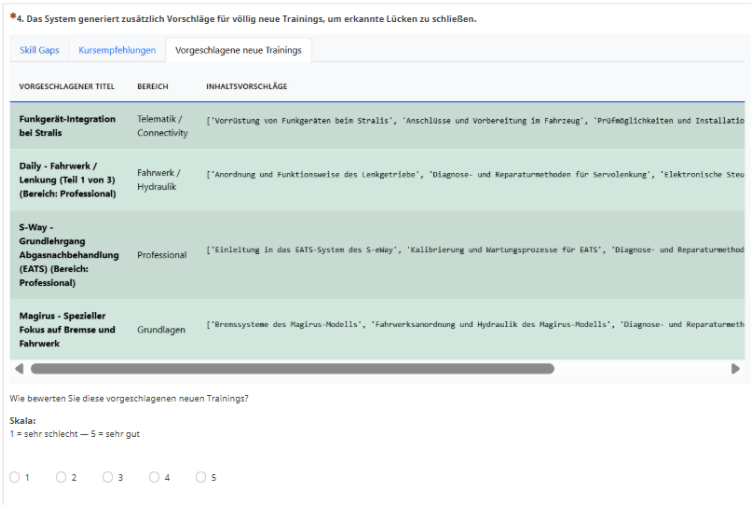
THEMA ID	EMPFOHLENER KURS	VERTRAUENSSCORE	BEGRÜNDUNG
1	**IVECO Services – Fahrzeugüberwachung und Verwaltung in eTim (Teil 5 von 5) (Bereich: Konnektivitätsberater)**	hoch	Dieses Training befasst sich mit der Fahrzeugüberwachung und -verwaltung, einschließlich Telematik-Tools und Ansichten. Es ist relevant für die Behandlung von Tickets, die Telematics betreffen.
2	**S-Way - Aufbaulehrgang Elektrik (Teil 2 von 3) (Bereich: Professional)**	hoch	Dieses Training befasst sich mit der Arbeit an elektronischen Systemen und bietet relevante Informationen zur Telematik.
4	**Grundlehrgang - Elektrik (Bereich: Service Mechaniker)**	hoch	Dieses Training beinhaltet grundlegende Informationen über die elektrischen Systeme in Fahrzeugen, einschließlich der Verständnis der Aufbau und Funktionsweise von Elektrik-Systemen.
5	**S-Way - Aufbaulehrgang Elektrik (Teil 1 von 3) (Bereich: Professional)**	hoch	Dieses Training beinhaltet Informationen zur Abgasnachbehandlung und könnte relevante Inhalte für die Bearbeitung der Tickets enthalten.
6	**EuroCargo - Grundlehrgang der Baureihe (Bereich: Grundlagen)**	mittel	Dieses Training bietet eine Übersicht über die Fahrzeugbaureihen, was den Eurotraiker auch abdeckt.
6	**Wissensstandsanalyse - Motor & Abgassystem (Bereich: Service Mechaniker)**	hoch	Dieses Training beinhaltet Tests und Quiz zu den Motoren und Abgassystemen, was den Eurotraiker abdeckt.
7	**Wissensstandsanalyse - Motor & Abgassystem (Bereich: Service Mechaniker)**	hoch	Dieses Training beinhaltet den Wissensstand zum Motor und Abgassystem, was die Probleme mit dem Antrieb abdeckt.
8	**Daily - Elektrik / Elektronik (Teil 3 von 3) (Bereich: Professional)**	hoch	Dieses Training beinhaltet Prüfmöglichkeiten am Fahrzeug, die relevant für Telematik-Updates und -diagnosen sind.
9	**Grundlehrgang - Elektrik (Bereich: Service Mechaniker)**	hoch	Dieses Training beinhaltet die Grundlagen der Elektronik und die Kenntnis von Datenbussystemen, was für das Verständnis von Fehlercodes hilfreich ist.
9	**Daily - Elektrik / Elektronik (Teil 3 von 3) (Bereich: Professional)**	hoch	Dieses Training fokussiert sich auf die Prüfmöglichkeiten am Fahrzeug, was für das Diagnose- und Fehlerbehandlungsvermögen relevant ist.
10	**Daily - Elektrik / Elektronik (Teil 1 von 3) (Bereich: Professional)**	hoch	Dieses Training beinhaltet Informationen zur Multiplex-Architektur und anderen elektronischen Systemen, die relevant für Fehlercodes und OTA-Updates sind.
11	**Grundwissen - Motor (Teil 1 von 2) (Bereich: Service Mechaniker)**	hoch	Dieses Training bietet einen Überblick über die Motorentechnik und kann Probleme mit dem Motor des Massif beheben.
11	**Grundwissen - Motor (Teil 2 von 2) (Bereich: Service Mechaniker)**	hoch	Dieses Training bietet fortgeschrittene Kenntnisse über die Motorentechnik und kann komplexe Probleme mit dem Motor des Massif beheben.
13	**Telematik / Connectivity - Grundlagen (Bereich: Professional)**	mittel	Dieses Training beinhaltet die Grundlagen der Telematik, was für das Verständnis und Support von Fahrzeugen mit Telematikkomponenten hilfreich ist.

Wie bewerten Sie die Qualität dieser Trainings-Empfehlungen?

Skala:
1 = sehr schlecht — 5 = sehr gut

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Figure 7.9: Part 2: Course Recommendations table (Q3)



*4. Das System generiert zusätzlich Vorschläge für völlig neue Trainings, um erkannte Lücken zu schließen.

Skill Gaps | Kursempfehlungen | Vorgeschlagene neue Trainings

VORGESCHLAGENER TITEL	BEREICH	INHALTSVORSCHLÄGE
Funkgerät-Integration bei Stralis	Telematik / Connectivity	['Vorrüstung von Funkgeräten beim Stralis', 'Anschlüsse und Vorbereitung im Fahrzeug', 'Prüfmöglichkeiten und Installatio
Daily - Fahrwerk / Lenkung (Teil 1 von 3) (Bereich: Professional)	Fahrwerk / Hydraulik	['Anordnung und Funktionsweise des Lenkgetriebe', 'Diagnose- und Reparaturmethoden für Servolenkung', 'Elektronische Steu
S-Way - Grundlehrgang Abgasnachbehandlung (EATS) (Bereich: Professional)	Professional	['Einleitung in das EATS-System des S-eWay', 'Kalibrierung und Wartungsprozesse für EATS', 'Diagnose- und Reparaturmethod
Magirus - Spezieller Fokus auf Bremse und Fahrwerk	Grundlagen	['Bremssysteme des Magirus-Modells', 'Fahrwerksanordnung und Hydraulik des Magirus-Modells', 'Diagnose- und Reparaturmeth

Wie bewerten Sie diese vorgeschlagenen neuen Trainings?

Skala:
1 = sehr schlecht — 5 = sehr gut

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Figure 7.10: Part 2: Proposed New Trainings table (Q4)

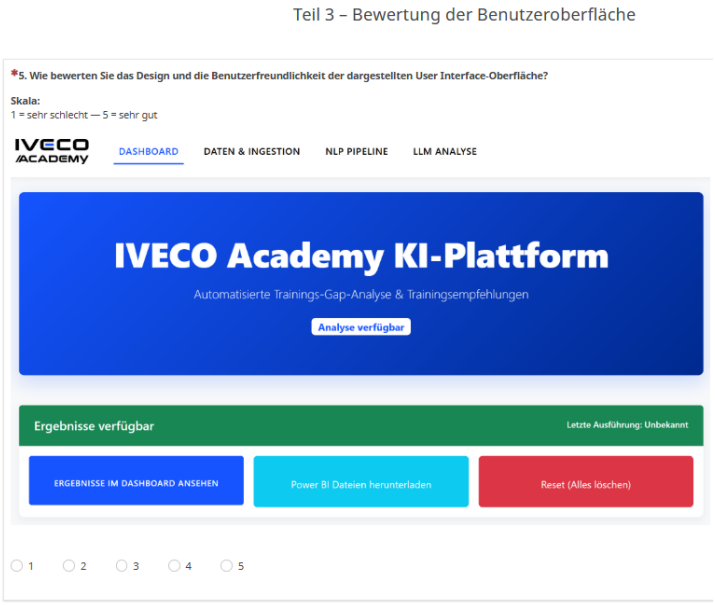


Figure 7.11: Part 3: Main Dashboard screenshot (Q5, UI/UX)

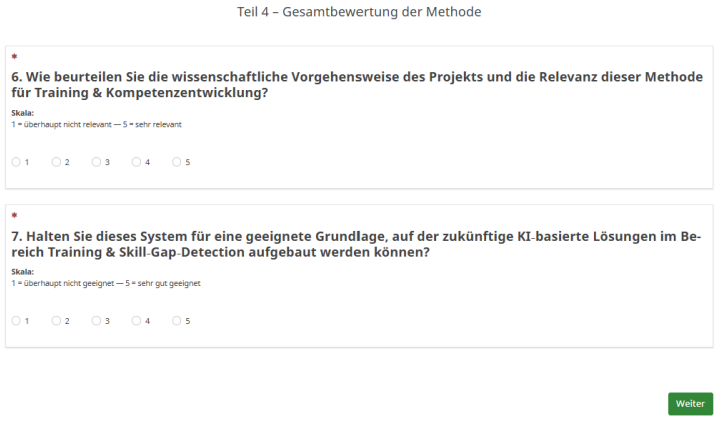


Figure 7.12: Part 4: Scientific approach relevance (Q6) and future AI foundation (Q7)

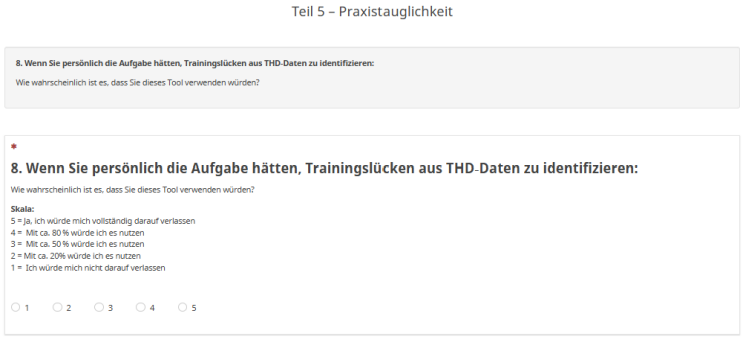


Figure 7.13: Part 5: Personal adoption likelihood (Q8) and improvement suggestions (Q9)

Appendix D: Declaration of Artificial Intelligence Tool Usage

In accordance with the academic integrity guidelines of Hochschule Neu-Ulm, the author declares that the following AI-based tools were used during the preparation of this thesis. All AI-assisted outputs were critically reviewed, verified, and substantially edited by the author. The intellectual responsibility for all content, arguments, and conclusions presented in this thesis remains solely with the author.

Tool	**Purpose of Use**
ChatGPT (v5.2)	Brainstorming of research ideas, prompt engineering for pipeline development, and evaluation of intermediate results.
Claude Sonnet (v4.5 / v4.6)	Enhancing academic writing quality, supporting the structuring and drafting of the literature review, and improving formal language consistency across chapters.
Gemini Pro (v3.1)	Generating and refining figures and visual diagrams used throughout the thesis.
Google NotebookLM	Summarising academic papers and supporting the organisation of literature review sources.
SciSpace	Discovery of relevant academic papers, literature review support, and exploration of related research domains.

Table 7.1: AI tools used during the preparation of this thesis and their respective purposes.

The author confirms that no AI tool was used to autonomously generate conclusions, fabricate data, or produce content that was incorporated without critical review and personal verification.

References

Agrawal, A., Fu, W., & Menzies, T. (2018). What is wrong with topic modeling? and how to fix it using search-based software engineering. *Information and Software Technology*, *98*, 74–88.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*, 993–1022.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernardo, M. S., Bernstein, M. S., Bhagat, A., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Campello, R. J., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In *17th pacific-asia conference on knowledge discovery and data mining (pakdd 2013)* (pp. 160–172). Springer.

Capaldo, G., Iandoli, L., & Zollo, G. (2006). A situationalist perspective to competency management. *Human Resource Management*, *45*(3), 429–448.

Cerasoli, C. P., Alliger, G. M., Donsbach, J. S., Mathieu, J. E., Tannenbaum, S. I., & Orvis, K. A. (2018). Antecedents and outcomes of informal learning behaviors: A meta-analysis. *Journal of Business and Psychology*, *33*(2), 203–230.

Chen, Y., et al. (2019). Intent classification and response generation in technical support dialogue. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, *36*, 10088–10115.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Edge, D., et al. (2024). From local to global: A graph rag approach to query-focused summarization. *arXiv preprint*.

Efstathiou, V., et al. (2018). Word embeddings for the software engineering domain. *Mining Software Repositories*.

Ferreira, R. R., & Abbad, G. (2013). Training needs assessment: Where we are and where we should go. *BAR-Brazilian Administration Review*, *10*, 77–99.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Goldstein, I. L., & Ford, J. K. (2002). *Training in organizations: Needs assessment, development, and evaluation*. Wadsworth/Thomson Learning.

Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS quarterly*, *37*(2), 337–355.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Gubbins, C., Harney, B., van der Werff, L., & Rousseau, D. M. (2018). Enhancing the trustworthiness and credibility of human resource development: Evidence-based management to the rescue? *Human Resource Development Quarterly*, *29*(3), 193–202.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS quarterly*, *28*(1), 75–105.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Ismail, A. (2016). Designing training programs based on needs assessment. *Journal of Workplace Learning*.

Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.

Kraiger, K. (2014). Looking back and looking forward: Trends in training and development research. *Human Resource Development Quarterly*, *25*(4), 401–408.

Lau, J. H., & Baldwin, T. (2016). An empirical evaluation of doc2vec with practical insights into document embedding generation. *Proceedings of the 1st workshop on representation learning for NLP*, 78–86.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474.

Lin, C., et al. (2002). Latent user needs inference. *Journal of Information Science*.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, *12*, 157–173. https://doi.org/10.1162/tacl_a_00638

Mallen, A., Asai, A., Zhong, V., Das, R., Hajishirzi, H., & Khashabi, D. (2023). When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9802–9822.

Manuti, A., Pastore, S., Scardigno, A. F., Giancaspro, M. L., & Morciano, D. (2015). Formal and informal learning in the workplace: A research review. *International Journal of Training and Development*, *19*(1), 1–17.

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision support systems*, *15*(4), 251–266.

Marsick, V. J., & Watkins, K. E. (2003). Demonstrating the value of an organization's learning culture: The dimensions of the learning organization questionnaire. *Advances in developing human resources*, *5*(2), 132–151.

McInnes, L., Healy, J., & Astels, S. (2017). Hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, *2*(11), 205.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.

Muriana, C., & Vizzini, G. (2017). Project risk management: A deterministic quantitative technique for assessment and mitigation. *International Journal of Project Management*, *35*(3), 320–340.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, 100–108.

Nguyen, T., et al. (2023). Technology-based knowledge sharing and employee performance: An empirical study. *Journal of Knowledge Management*.

Noe, R. A. (2020). *Employee training and development*. McGraw-Hill Education.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, *24*(3), 45–77.

Peng, B., Li, C., He, P., Galley, M., & Gao, J. (2023). Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Pham, T., et al. (2019). Training needs assessment in modern organizations. *International Journal of Human Resources*.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *Proceedings of the eighth ACM international conference on Web search and data mining*, 399–408.

Rosen, C., & Shihab, E. (2016). What are mobile developers asking about? a large scale study using stack overflow. *Empirical Software Engineering*, *21*(3), 1192–1223.

Sallam, M. (2023). Chatgpt utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare*, *11*(6), 887.

Schick, T., Dwivedi-Yu, J., Dessi, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*.

Shen, Y., Song, K., Tan, X., Li, D., Lu, W., & Zhuang, Y. (2023). HuggingGPT: Solving AI tasks with ChatGPT and its friends in Hugging Face. *Advances in Neural Information Processing Systems*, *36*, 37517–37530.

Shi, W., Min, S., Michalski, M., Sil, A., Hajishirzi, H., & Zettlemoyer, L. (2023). Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Simon, H. A. (1996). *The sciences of the artificial*. MIT press.

Sun, J., Cheng, C., Fang, H., et al. (2023). Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. *arXiv preprint arXiv:2307.07697*.

Tian, Y., et al. (2015). Automated routing of it support tickets. *Mining Software Repositories*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: A framework for evaluation in design science research. *European journal of information systems*, *25*(1), 77–89.

vom Brocke, J., Hevner, A. R., & Maedche, A. (2020). *Design science research cases*. Springer.

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing theory-driven user-centric explainable ai. *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–15.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.

Wieringa, R. J. (2014). *Design science methodology for information systems and software engineering*. Springer.

World Economic Forum. (2023). *Future of jobs report 2023* (tech. rep.). World Economic Forum.

Zangari, M., et al. (2023). Ticket automation in it support: A comprehensive review. *Expert Systems with Applications*.

Zanker, M., et al. (2019). Recommender systems and user needs. *User Modeling and User-Adapted Interaction*.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.