NARRATE

Providing trustful and ethical personalised conversational interfaces on top of news and information

DELIVERABLE D3.2
SotA LLMs

Project number: ITEA 23036
Document version no.: v0.9
Edited by: Yunus Akyol
Date: 06.01.2025

ITEA Roadmap challenge:
Smart Communities

HISTORY

Document version #	Date	Remarks
V0.1	03.11.2025	Starting version, template
V0.2		Compilation of first input by partners
V0.3		Second compilation of input by partners
V0.5	19.12.2025	First version with almost all the contents
V0.9	06.01.2026	Final review and administrative completion
V1.0		Final version

Contents

1. Executive Summary

Large Language Models (LLMs) have rapidly evolved over the past few years, reaching a level of maturity that enables their adoption across a wide range of enterprise domains, including Human Resources (HR). Recent advances in transformer-based architectures, large-scale pre-training, instruction tuning, and retrieval-augmented generation (RAG) have significantly improved the reasoning, language understanding, and task generalization capabilities of these models. Both commercial and open-source LLM ecosystems now offer scalable, customizable, and domain-adaptable solutions suitable for complex organizational workflows.

In the HR domain, LLMs are increasingly applied to recruitment and selection, onboarding and training, knowledge management, employee support, and experience analytics. These applications promise enhanced efficiency, improved decision support, and more personalized employee interactions. However, the deployment of LLMs in HR contexts also introduces critical challenges related to bias and fairness, explainability, data privacy, regulatory compliance, and system integration with existing enterprise infrastructures.

This report provides a state-of-the-art overview of Large Language Models with a specific focus on their relevance to HR processes. It reviews recent developments in LLM architectures, training and evaluation methodologies, and responsible AI practices, and examines how these advances translate into HR-focused use cases. Furthermore, it identifies current research gaps and technical challenges that must be addressed to ensure trustworthy, effective, and compliant LLM-based HR solutions.

The findings of this report directly inform the objectives of the Narrate project by establishing a clear technological baseline, highlighting innovation opportunities, and supporting the design of LLM-enabled HR use cases that are robust, ethical, and aligned with real-world organizational needs.

2. Introduction

2.1 Background and Motivation

Over the past decade, advances in Natural Language Processing (NLP) have been largely driven by the emergence of LLMs, which leverage transformer-based architectures and large-scale data-driven learning. Models such as GPT, PaLM, LLaMA, Claude, and Gemini have demonstrated unprecedented performance across a broad spectrum of language understanding and generation tasks, including reasoning, summarization, question answering, and dialogue. These capabilities have significantly expanded the applicability of AI systems beyond narrow task-specific solutions toward more general-purpose, adaptable language technologies (Brown et al., 2020; Vaswani et al., 2017; Touvron et al., 2023).

In parallel with these technological advancements, organizations are increasingly seeking intelligent systems capable of supporting complex, knowledge-intensive business processes. HR functions represent a particularly relevant domain, as they rely heavily on textual data, human-centric decision-making, and continuous interaction with employees and candidates. HR processes such as recruitment, onboarding, learning and development, policy management, and employee support involve large volumes of unstructured information, making them well-suited for LLM-based approaches.

Recent progress in instruction tuning, reinforcement learning from human feedback (RLHF), and RAG has further enhanced the reliability and domain adaptability of LLMs in enterprise settings (Ouyang et al., 2022; Lewis et al., 2020). These techniques enable models to align more closely with organizational goals, incorporate proprietary knowledge sources and reduce hallucinations, which are all key requirements for HR-related applications where accuracy, fairness and trust are critical.

At the same time, the deployment of LLMs in HR contexts raises important challenges and risks. Issues related to algorithmic bias, transparency, explainability, data privacy, and regulatory compliance, particularly under frameworks such as the EU AI Act, GDPR and KVKK must be carefully addressed to ensure responsible adoption. Consequently, a clear understanding of the current state of the art, as well as its limitations, is essential for designing HR-focused LLM solutions that are both effective and ethically sound.

Against this background, this report is motivated by the need to systematically assess recent developments in LLM research and development, with a specific focus on their applicability to HR processes. By synthesizing technical advances, evaluation practices, and real-world use cases, the report aims to provide a solid foundation for informed decision-making and innovation within the scope of the Narrate project.

2.2 Objectives of the Report

The primary objective of this report is to provide a comprehensive and up-to-date overview of the state of the art in LLMs, with a particular focus on their applicability to HR processes. The report aims to synthesize recent advances in LLM architectures, training paradigms, evaluation methodologies, and responsible AI practices, drawing on both academic research and industrial developments.

More specifically, the objectives of the report are to:

- Establish a **technical baseline** for understanding contemporary LLM capabilities, limitations, and trends relevant to enterprise deployment.
- Analyze how state-of-the-art LLM techniques translate into **practical HR applications**, including recruitment, onboarding, knowledge management, and employee experience.
- Identify **key challenges and research gaps** that remain unresolved in the context of HR-oriented LLM systems, such as domain adaptation, bias mitigation, explainability, and data privacy.

- Provide a **structured reference framework** to support informed design decisions within the Narrate project, ensuring alignment with technological best practices and regulatory expectations.
- Highlight **innovation opportunities** where LLM-based approaches can offer measurable improvements over traditional HR information systems.

By addressing these objectives, the report is intended to serve both as a technical reference and as a strategic input for the Narrate project, supporting the development of robust, trustworthy, and scalable LLM-enabled HR solutions.

2.3 Scope and Limitations

This report examines the state of the art in LLMs with a focus on their application to HR processes. The analysis covers recent advances in LLM architectures, training and fine-tuning strategies, evaluation practices, and responsible AI considerations. Both commercial and open-source models are included to provide a balanced perspective.

The HR-related scope includes use cases such as recruitment and selection, onboarding and training, knowledge management, employee support, and employee experience analytics. The report concentrates on text-based and conversational applications, where LLMs are currently most mature. Multimodal models are considered only when directly relevant to HR scenarios.

The report has several limitations. Due to the fast-evolving nature of the field, the findings represent a snapshot rather than a definitive or exhaustive assessment. Detailed model-to-model comparisons, large-scale empirical evaluations on proprietary HR datasets, and in-depth legal or organizational analyses are beyond the scope of this document. Implementation-level concerns, including deployment architectures and cost optimization, are addressed only where they directly affect the Narrate project objectives and do not include confidential information.

2.4 Document Structure

This document is organized as follows. Section 3 provides background information on LLMs, including their evolution, underlying architectures, and key training paradigms. Section 4 reviews the current state of the art in LLM research and development, covering model architectures, training techniques, evaluation methodologies, and responsible AI considerations.

Section 5 examines the application of LLMs within the Human Resources domain, focusing on key HR processes such as recruitment, onboarding, knowledge management, and employee experience. Section 6 identifies open research gaps and challenges related to domain adaptation, fairness, explainability, privacy, and system integration.

Section 7 relates the state-of-the-art findings to the objectives of the Narrate project, mapping technical capabilities to concrete use cases and expected innovations. Finally, Section 8 summarizes the main conclusions and discusses the readiness of LLM technologies for HR applications, followed by references and supplementary material in Sections 9 and 10.

3. Background on Large Language Models

3.1 Evolution of NLP and LLMs

NLP has evolved significantly over the past decades, transitioning from rule-based and statistical methods to data-driven neural approaches. Early NLP systems relied on hand-crafted rules and symbolic representations, which were limited in scalability and robustness. The introduction of statistical models and probabilistic methods, such as n-gram language models and hidden Markov models, enabled more flexible handling of linguistic variability but still struggled with long-range dependencies and semantic understanding (Jurafsky & Martin, 2009).

The adoption of deep learning marked a major turning point in NLP. Recurrent neural networks (RNNs) and their variants, including long short-term memory (LSTM) and gated recurrent unit (GRU) models, improved sequence modeling by capturing contextual information across time steps. However, these architectures faced challenges related to parallelization, training efficiency, and the modeling of very long contexts (Hochreiter & Schmidhuber, 1997).

A fundamental shift occurred with the introduction of the transformer architecture, which replaced recurrent computation with self-attention mechanisms (Vaswani et al., 2017). Transformers enabled efficient parallel training and more effective modeling of long-range dependencies, laying the foundation for large-scale pre-trained language models. This paradigm facilitated the emergence of models such as BERT, GPT, and their successors, which demonstrated that pre-training on massive text corpora followed by task-specific fine-tuning could yield strong performance across diverse NLP tasks (Devlin et al., 2019; Radford et al., 2019).

More recently, the scaling of model size, training data, and computational resources has led to the development of Large Language Models capable of few-shot and zero-shot generalization. Models such as GPT-3, PaLM, LLaMA, and subsequent generations have shown that increasing scale, combined with improved training objectives and alignment techniques, results in emergent capabilities such as reasoning, instruction following, and contextual adaptation (Brown et al., 2020; Chowdhery et al., 2022; Touvron et al., 2023). These advances have positioned LLMs as general-purpose language engines, enabling their adoption in complex enterprise domains, including Human Resources.

3.2 Transformer Architecture Overview

The transformer architecture has become the foundational building block of modern Large Language Models due to its ability to efficiently model long-range dependencies and scale to very large datasets. Introduced by Vaswani et al. (2017), the transformer replaces recurrent and convolutional structures with a self-attention mechanism that allows each token in a sequence to attend to all other tokens in parallel. This design significantly improves training efficiency and enables effective utilization of modern hardware accelerators.

At the core of the transformer is the **self-attention mechanism**, which computes contextualized representations by weighting the relevance of each token with respect to others in the sequence. Multi-head attention extends this concept by allowing the model to capture different types of relationships such as syntactic, semantic, or positional dependencies within the same layer. Positional encodings are introduced to retain information about token order, which is otherwise absent in the attention-based formulation.

Transformers are typically composed of stacked layers that alternate between attention blocks and position-wise feed-forward networks, combined with residual connections and normalization techniques to stabilize training. Decoder-only transformer variants, which predict the next token autoregressively, form the basis of many prominent LLMs, while encoder-only and encoder–decoder variants are used for tasks such as representation learning and sequence-to-sequence generation.

The architectural properties of transformers enable favorable **scaling behavior**, where increases in model size, data volume, and compute tend to yield predictable performance gains. This observation has motivated large-scale training regimes and informed the design of contemporary LLM families (Kaplan et al., 2020; Hoffmann et al., 2022). For enterprise applications, including HR systems, transformer-based LLMs offer a flexible and extensible architecture that can be adapted through fine-tuning or augmentation to support diverse language-centric workflows.

3.3 Pre-Training and Fine-Tuning Paradigms

Modern LLMs are typically developed through a two-stage learning process consisting of large-scale pre-training followed by task or domain-specific adaptation. During pre-training, models are trained on massive, diverse text corpora using self-supervised objectives, most commonly next-token prediction. This phase enables LLMs to acquire broad linguistic knowledge, general world understanding, and latent reasoning capabilities without reliance on labeled data (Brown et al., 2020).

Fine-tuning strategies are subsequently applied to adapt pre-trained models to specific tasks, domains, or interaction styles. Early approaches relied on supervised fine-tuning using labeled datasets, while more recent methods emphasize **instruction tuning**,

where models are trained on collections of natural-language instructions paired with desired outputs. Instruction tuning has been shown to significantly improve model usability, generalization, and alignment with human expectations, particularly in conversational and enterprise contexts (Wei et al., 2022).

To further align model behavior with human values and reduce undesirable outputs, **RLHF** has become a standard component of LLM training pipelines. RLHF involves training a reward model based on human preferences and optimizing the LLM to maximize this reward, resulting in improved helpfulness, safety, and controllability (Ouyang et al., 2022). This paradigm is especially relevant for HR applications, where inappropriate or biased responses can have significant organizational and legal implications.

In parallel, **parameter-efficient fine-tuning (PEFT)** methods, such as adapters, prefix tuning, and low-rank adaptation (LoRA), have emerged as practical alternatives to full model fine-tuning. These techniques enable domain adaptation with substantially reduced computational and data requirements, making them attractive for enterprise environments with limited resources or strict data governance constraints (Hu et al., 2022).

Together, these pre-training and fine-tuning paradigms provide a flexible toolkit for adapting LLMs to HR-specific tasks, enabling organizations to leverage general-purpose models while incorporating domain knowledge, organizational policies, and ethical constraints.

3.4 Key LLM Families and Ecosystem

The current LLM landscape comprises a diverse ecosystem of commercial and open-source model families, each characterized by different design choices, training strategies, and deployment models. Commercial offerings such as OpenAI's GPT series, Anthropic's Claude models, and Google's Gemini family represent highly optimized, large-scale systems trained on extensive proprietary datasets. These models typically emphasize strong general reasoning capabilities, instruction following, and safety alignment, making them well-suited for enterprise-grade applications, including HR support systems and conversational assistants.

In parallel, open-source LLM families have gained significant momentum, offering greater transparency, customization, and on-premise deployment options. Meta's LLaMA models and their successors have become foundational within the open-source community, enabling fine-tuning and specialization for domain-specific use cases. Similarly, models such as Mistral and Mixtral have introduced architectural innovations, including mixture-of-experts (MoE) designs, which improve computational efficiency while maintaining competitive performance (Touvron et al., 2023; Jiang et al., 2023).

The broader LLM ecosystem extends beyond base models to include tooling and infrastructure for model adaptation, orchestration, and integration. Frameworks for

prompt engineering, fine-tuning, retrieval augmentation, and monitoring have become integral components of enterprise LLM deployments. In HR contexts, this ecosystem supports the development of tailored solutions that integrate organizational knowledge bases, comply with data governance requirements, and align model behavior with corporate policies.

From an application perspective, the availability of both commercial and open-source LLMs enables organizations to balance performance, control, and compliance considerations. While commercial models often provide state-of-the-art performance and managed services, open-source alternatives offer flexibility and data sovereignty, which are particularly relevant in regulated domains such as HR.

Table 1: *Comparative Performance of State-of-the-Art LLM Families (Hui et al., 2025)*

Model	**Type**	**MMLU (Knowledge)**	**HumanEval (Coding)**	**GSM8K (Math)**	**Context Window**
GPT-4o	Proprietary	88.7%	90.2%	96.1%	128k
Claude 3.5 Sonnet	Proprietary	88.3%	92.0%	96.4%	200k
Gemini 1.5 Pro	Proprietary	85.9%	84.7%	94.0%	2M+
Llama 3.1 (405B)	Open Weights	88.6%	89.0%	96.8%	128k
Qwen 2.5 (72B)	Open Weights	86.8%	88.2%	96.6%	128k
Mistral Large 2	Proprietary/Open	84.0%	92.2%	93.3%	128k

Benchmark Descriptions:

- MMLU (Massive Multitask Language Understanding): Measures general world knowledge and problem-solving ability across 57 subjects. Critical for understanding HR domains and regulations.
- HumanEval: Measures code generation capability. Essential for converting natural language queries into structured database queries (SQL/JSON) in RAG systems.
- GSM8K: Measures multi-step mathematical reasoning. Crucial for accurate calculations of salary, leave balances, and compensation analytics.

As demonstrated in the table, open-weight models such as Llama 3.1 and Qwen 2.5 have reached performance levels comparable to proprietary models like GPT-4o and Claude 3.5 Sonnet in critical reasoning tasks. This parity allows the project to utilize on-

premises open-source models without compromising intelligence or accuracy, thereby ensuring data privacy compliance.

3.5 RAG and Augmented Models

Retrieval-Augmented Generation (RAG) has emerged as a key technique for extending the capabilities of Large Language Models by combining parametric language generation with non-parametric information retrieval. Instead of relying solely on knowledge encoded during pre-training, RAG-based systems dynamically retrieve relevant documents from external knowledge sources and condition the model's responses on this retrieved context, as defined in the foundational NeurIPS work by Lewis et al. (2020). This approach significantly improves factual accuracy, reduces hallucinations, and enables the incorporation of up-to-date or organization-specific information.

RAG architectures typically consist of three components: a retriever that selects relevant documents or passages, a knowledge store such as a vector database, and a generator that produces responses grounded in the retrieved content. Advances in embedding models and similarity search have further enhanced retrieval quality, making RAG a practical and scalable solution for enterprise use cases.

Beyond classical RAG, a broader class of augmented LLMs has emerged, incorporating tools, APIs, and structured data sources into the inference loop. These models can perform actions such as database queries, document analysis, and workflow orchestration, effectively functioning as intelligent agents. Such augmentation is particularly relevant for HR applications, where accurate access to internal policies, employee handbooks, training materials, and regulatory documents is essential. For HR systems, RAG-based approaches enable secure and controllable use of LLMs by grounding responses in verified organizational knowledge, thereby supporting trust, compliance, and explainability in decision-support and employee-facing applications.

In practice, relying solely on vector-based search can be insufficient when dealing with HR-specific terminology such as procedure codes or employee identifiers. The BEIR Benchmark demonstrates that vector models may experience performance degradation in zero-shot scenarios involving domain-specific terms not present in their training data (Thakur et al., 2021). To address this limitation, modern systems adopt hybrid search architectures that combine the precision of BM25 keyword-based retrieval with semantic vector search. Research by Microsoft shows that combining hybrid search with semantic reranking increases retrieval success from 43.8 percent to 60.1 percent compared to vector search alone (Microsoft, 2023), while AWS reports that hybrid approaches help prevent loss of retrieval precision (Amazon Web Services, 2023).

Dataset	BeIR/fiqa		
Method\Metric	Recall@1	Recall@4	Recall@10
bm25	0.112	0.215	0.297
dense	0.156	0.316	0.398
sparse	0.196	0.334	0.438
hybird_dense_sparse	0.203	0.362	0.456
hybird_dense_bm25	0.156	0.316	0.394

Figure 1 Hybrid Search Performance Comparison

4. Current State of the Art in LLM Research and Development

4.1 Model Architectures and Efficiency

Recent state-of-the-art developments in Large Language Models have increasingly focused not only on improving model capability but also on computational efficiency, scalability, and deployability. While early advances were largely driven by scaling model parameters, training data, and compute, more recent research emphasizes architectural innovations that achieve improved performance–efficiency trade-offs (Kaplan et al., 2020; Hoffmann et al., 2022). This shift is particularly relevant for enterprise HR applications, where cost, latency, reliability, and scalability are critical constraints.

One prominent architectural trend is the adoption of Mixture-of-Experts (MoE) models, in which only a subset of model parameters is activated for each input. This design enables models to scale to very large parameter counts while keeping inference costs manageable. Architectures such as Switch Transformer and Mixtral demonstrate that MoE models can achieve competitive or superior performance compared to dense models at a fraction of the computational cost (Fedus et al., 2022; Jiang et al., 2023; Mistral AI, 2024).

In parallel, substantial progress has been made in model compression and optimization techniques. Quantization reduces numerical precision to enable faster inference and lower memory usage with minimal performance degradation, while knowledge distillation transfers capabilities from large teacher models to smaller student models suitable for resource-constrained deployments. Sparse modeling and structured pruning further reduce computational overhead by selectively disabling redundant parameters (Frantar et al., 2022).

Another important development is the expansion of context length and refinement of memory mechanisms, allowing LLMs to process longer documents and sustained interactions. Extended context windows and attention optimizations support enterprise

use cases such as HR policy analysis, contract review, and longitudinal employee support workflows.

At the architectural level, the industry has largely converged on decoder-only transformer models, such as Llama 3 and Qwen 2.5, due to their strong performance in generative tasks and in-context learning compared to encoder–decoder alternatives (Dubey et al., 2024). In parallel, LLM architectures are evolving to handle non-textual enterprise data common in HR processes. Vision-Language Models such as ColPali and SigLIP enable direct processing of visually structured documents, including tables and charts, without reliance on OCR, thereby preserving layout and contextual information (Faysse et al., 2024).

Overall, the current state of the art reflects a clear transition from scale-driven progress toward architectures that balance performance, efficiency, and controllability. This evolution is essential for the sustainable and responsible adoption of LLMs in enterprise HR systems.

4.2 Training Techniques

State-of-the-art Large Language Models increasingly rely on sophisticated training techniques to improve performance, generalization, and alignment while managing the high costs of large-scale learning. Beyond raw data and compute scaling, recent advances emphasize data quality, training efficiency, and post-training alignment strategies, which are particularly important for sensitive enterprise domains such as Human Resources.

A key trend is the growing focus on data curation and filtering. High-quality, diverse, and well-balanced training data has been shown to be as important as model scale for achieving strong performance. Techniques such as deduplication, toxicity filtering, and domain balancing are now standard practice in large-scale LLM training pipelines, helping to reduce noise, bias, and undesirable behaviors (Raffel et al., 2020; Hoffmann et al., 2022). Curriculum learning and staged training approaches further support this goal by controlling the order and difficulty of training samples, enabling models to acquire foundational language capabilities before progressing to more complex reasoning and instruction-following tasks.

Post-training alignment remains a central component of modern LLM development. Reinforcement Learning from Human Feedback and related preference optimization methods are widely used to align model outputs with human expectations and safety constraints. More recent approaches such as Direct Preference Optimization aim to simplify and stabilize alignment by directly optimizing preference-based objectives, reducing reliance on complex reinforcement learning pipelines (Rafailov et al., 2023).

For domains like HR, retraining large models from scratch is impractical, which has led to increased emphasis on efficient adaptation techniques. Instruction tuning aligns

models with user intent by fine-tuning on instruction–response pairs, enabling reliable responses to structured HR queries such as policy explanations (Wei et al., 2021). In parallel, parameter-efficient fine-tuning methods such as Low-Rank Adaptation have become the standard for domain specialization. By freezing pre-trained weights and introducing a small number of trainable parameters, LoRA enables adaptation to internal HR terminology and policies with minimal computational overhead, reduced data exposure, and lower operational costs (Hu et al., 2022).

Collectively, these training and adaptation strategies reflect a shift toward more controlled, data-efficient, and adaptable LLM development practices, supporting reliable and compliant deployment in evolving enterprise HR environments.

4.3 Evaluation Methodologies

Evaluating Large Language Models is inherently challenging due to their broad, open-ended capabilities, which are not fully captured by traditional task-specific metrics. As a result, state-of-the-art evaluation has shifted toward multi-dimensional frameworks that combine standardized capability benchmarks, human preference assessments, and safety and robustness analyses.

In professional and enterprise settings, evaluation must extend beyond general academic tests to include domain-specific and retrieval-aware metrics. MMLU (Massive Multitask Language Understanding) has become a standard reference for measuring general knowledge and reasoning across diverse subject areas (Hendrycks et al., 2021). However, such multiple-choice benchmarks primarily assess static knowledge and exam-style reasoning, and may not reflect real-world interactive performance, tool use, or robustness under distribution shift.

For retrieval-augmented systems, the quality of retrieved context is as critical as answer correctness. Frameworks such as RAGAS evaluate Answer Relevancy, Context Relevancy, and Context Recall, providing a structured way to assess grounding and information access control. Recent studies demonstrate the effectiveness of these metrics in ensuring that unauthorized or sensitive data is filtered during retrieval-based interactions (Chen et al., 2025).

To address the need for broader coverage and transparency, holistic evaluation frameworks such as HELM propose multi-scenario and multi-metric assessments that include not only accuracy, but also calibration, robustness, fairness and bias, toxicity, and efficiency. This perspective is particularly relevant for HR applications, where risks related to biased or harmful outputs must be evaluated alongside usefulness (Liang et al., 2022).

Because many HR-relevant tasks are inherently open-ended, including policy explanation, CV summarization, and candidate communication, evaluation increasingly relies on human preference judgments and LLM-as-a-judge methodologies. Benchmarks such as MT-Bench and platforms like Chatbot Arena approximate human preferences through pairwise comparisons and automated judging, while also highlighting known limitations such as positional and verbosity biases that require mitigation (Zheng et al., 2023; LMSYS, 2023).

Overall, state-of-the-art evaluation emphasizes a combination of capability benchmarks, holistic multi-metric assessment, and preference-based interactive evaluation. For HR-focused systems, best practice further requires domain- and risk-aware testing, including HR-specific scenarios related to fairness, policy compliance, and sensitive data handling, as general benchmark improvements do not automatically translate into safe and compliant enterprise deployment.

4.4 Responsible and Ethical AI

As Large Language Models are increasingly deployed in real-world and high-stakes contexts, responsible and ethical AI practices have become a central component of state-of-the-art LLM development. This is particularly critical for Human Resources applications, where AI-supported systems may influence hiring decisions, employee evaluation, and access to organizational opportunities.

One of the most prominent ethical challenges is bias and fairness. LLMs trained on large-scale web data may inherit and amplify societal biases related to gender, ethnicity, age, or socioeconomic status. State-of-the-art mitigation strategies include dataset auditing and filtering, bias-aware fine-tuning, controlled prompting, and post-hoc output moderation. More recent approaches, such as Self-RAG and counterfactual data augmentation, enable models to critique and adjust their own outputs for potential bias before final response generation (Bender et al., 2021; Mehrabi et al., 2021; Asai et al., 2024). Nevertheless, bias cannot be fully eliminated at the model level alone, leading HR deployments to emphasize system-level safeguards such as human oversight and continuous bias monitoring.

Transparency and explainability represent another critical dimension of responsible AI. Due to their scale and complexity, LLMs are often perceived as opaque systems, which complicates trust and accountability. In HR contexts, black-box outputs are unacceptable, particularly in candidate screening and evaluation scenarios. Retrieval-Augmented Generation architectures support explainability by grounding responses in verifiable source documents and enabling citation-based generation, while additional practices such as prompt traceability and structured outputs further enhance transparency and regulatory compliance (Lewis et al., 2020).

Privacy and data protection are equally essential, as HR systems routinely process sensitive personal data. Compliance with frameworks such as GDPR and KVKK requires data minimization, strict access control, and careful deployment choices, including on-premises or private-cloud infrastructures. Parameter-efficient fine-tuning and retrieval-based architectures reduce the need to embed sensitive information directly into model weights, thereby lowering privacy risks during both training and inference.

Finally, responsible LLM deployment increasingly incorporates governance and lifecycle management practices. These include clear documentation of model capabilities and limitations, continuous monitoring for harmful behaviors, and regular updates aligned with evolving organizational policies and regulatory requirements. Emerging frameworks

such as the EU AI Act further reinforce the need for risk-based assessment and accountability, particularly for AI systems used in employment-related contexts.

Overall, the state of the art in responsible and ethical AI emphasizes a holistic approach that combines technical safeguards, organizational processes, and regulatory alignment. For HR applications, these considerations are core design requirements that directly determine the trustworthiness and feasibility of LLM-based solutions.

5. LLM Applications in the HR Domain

LLMs are increasingly applied across Human Resources (HR) functions, enabling scalable analysis of unstructured text and more natural interaction with HR knowledge and workflows. Compared to traditional rule-based or keyword-driven systems, LLMs support semantic understanding, contextual reasoning, and flexible language generation, which makes them particularly suitable for recruitment screening, onboarding support, HR knowledge management, and employee experience analytics. At the same time, state-of-the-art enterprise deployments generally position LLMs as decision-support tools embedded in governance frameworks rather than autonomous decision-makers.

5.1 Recruitment and Selection

Recruitment and selection represent one of the most mature and widely explored application areas for LLMs in HR. These processes involve intensive interaction with unstructured textual data, including job descriptions, curricula vitae (CVs), cover letters, interview transcripts, and candidate communications, making them amenable to LLM-based automation and decision support.

A prominent application is **CV and profile analysis**, where LLMs extract, normalize, and summarize candidate information. In contrast to keyword-based screening, LLMs can capture semantic similarities between job requirements and candidate experience, reducing dependence on exact term matching and improving recall. Recent approaches combine LLMs with embedding-based retrieval to support ranking, shortlisting, and candidate–role matching while enabling configurable screening criteria (Bhatia et al., 2023; Zhang et al., 2024).

LLMs are also increasingly employed for **job description generation and optimization**. By analyzing organizational competency frameworks and prior postings, models can draft inclusive, role-specific descriptions and suggest language modifications that reduce unintentional bias, supporting consistency and alignment with diversity and inclusion objectives.

In the interview stage, LLMs contribute to **interview preparation and analysis**, including generating structured interview questions, competency-based evaluation templates, and post-interview summaries. When used as decision-support tools, they

can improve standardization and reduce interviewer variability, while best practice retains final hiring decisions under human control due to ethical and legal considerations.

Finally, **conversational recruitment assistants** (chatbots) are widely used in large-scale hiring contexts for candidate Q&A, pre-screening questionnaires, and scheduling, improving responsiveness and reducing administrative workload. Instruction tuning and RAG increasingly support policy-aligned responses and integration with applicant tracking systems (ATS) and internal knowledge sources.

Despite these benefits, recruitment remains a **high-risk domain** due to concerns around bias, fairness, and explainability. As a result, state-of-the-art systems increasingly incorporate auditing mechanisms, bias monitoring, and human-in-the-loop review workflows to ensure compliant and responsible use.

5.2 Onboarding and Training

LLMs are increasingly applied to onboarding and training processes to enable more personalized, scalable, and adaptive learning experiences. Traditional onboarding programs often rely on static documentation and standardized training modules that may not reflect individual roles, backgrounds, or learning needs. LLM-driven systems enable a shift toward interactive, context-aware onboarding support.

A common application is **LLM-powered onboarding assistants** that provide new employees with conversational access to policies, procedures, and role-specific information. When implemented using retrieval-augmented generation, these assistants can provide up-to-date, document-grounded responses, improving early engagement and reducing reliance on manual HR support.

In training and development, LLMs support **content generation and adaptation**, including training summaries, explanatory materials, and knowledge checks. Models can tailor content to roles, skill levels, and learning objectives, enabling just-in-time learning and continuous upskilling, particularly in organizations facing rapid technological or regulatory change.

LLMs also enable **skill gap analysis and learning pathway recommendation** by analyzing job requirements, employee profiles, and (where appropriate) performance-related data. State-of-the-art implementations emphasize transparency and human oversight to ensure recommendations are interpretable and aligned with organizational goals.

An emerging direction is the integration of LLMs with **Extended Reality (XR)** for immersive onboarding and performance support. Recent studies demonstrate that LLM-driven XR systems can support more natural interaction through “intelligent digital humans,” while multimodal XR assistants can provide real-time procedural guidance through AR overlays and context-aware feedback (Song & Xiong, 2025; Srinidhi et al.,

2024; Zhu et al., 2025). Research also suggests that XR training frameworks with adaptive feedback can improve motivation and reduce errors in procedural learning scenarios (Gianni et al., 2025).

Overall, LLM-enabled onboarding and training systems offer a pathway to more responsive learning ecosystems; however, accuracy, bias, and privacy considerations remain essential, especially when employee development data is involved.

5.3 Knowledge Management and HR Support

Knowledge management and HR support represent one of the most immediately deployable and impactful application areas for Large Language Models. HR departments typically manage extensive repositories of policies, procedures, benefits information, regulatory guidelines, and internal communications, which are often difficult to navigate efficiently using traditional search systems.

LLM-powered HR knowledge assistants, particularly when implemented using Retrieval-Augmented Generation, provide employees and managers with natural language access to verified internal documents. By retrieving and synthesizing information from authoritative sources, these systems deliver accurate and context-aware responses while maintaining traceability to underlying content. This capability reduces response times, alleviates HR helpdesk workload, and supports controlled grounding, auditability, and explainability, which are critical in HR contexts where incorrect guidance may carry legal or financial risk.

Beyond employee-facing support, LLMs can assist HR professionals by summarizing internal documentation, supporting policy drafting, and facilitating compliance-related workflows. RAG-based architectures further enable organizations to retain data ownership while minimizing hallucination risk, reinforcing trust in HR decision-support applications.

Evidence from adjacent domains indicates that grounding conversational agents in verified documents improves usability and user trust, and that hybrid interaction paradigms, such as conversational responses complemented by visual menus or text overlays, enhance user confidence and retention. These findings are particularly relevant for multimodal HR assistant deployments (Boumans et al., 2025; Cordioli et al., 2025).

To support real-time HR knowledge access, high-performance infrastructure is required. Vector database technologies play a central role in meeting low-latency and scalability requirements. Modern systems such as Qdrant and Weaviate demonstrate significantly higher throughput compared to traditional Elasticsearch-based infrastructures, with reported latencies below 10 milliseconds, which is critical for maintaining a responsive employee experience (Qdrant, 2024).

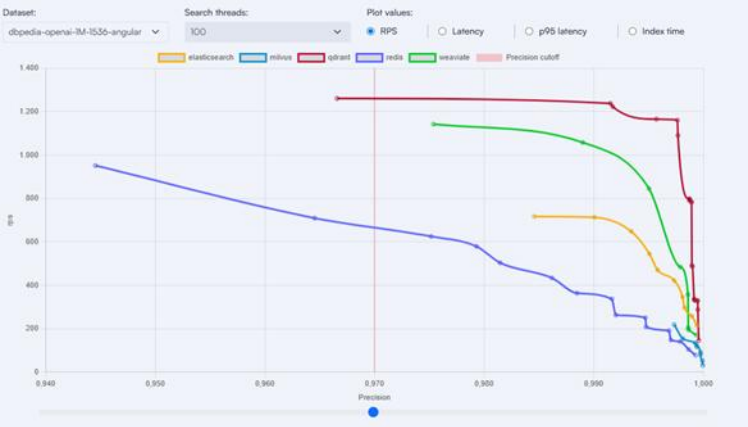

Figure 2 RPS Values of Vector DB

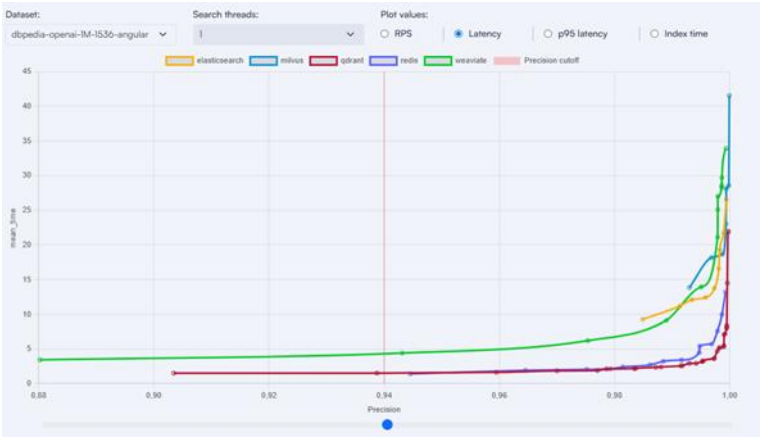

Figure 3 Latency Values of Vector DB

According to the Latenode report, the possession of Hybrid Search and Metadata Filtering capabilities by the database is a determining factor for success in corporate domains like HR. These features ensure that sensitive documents, such as "Salary Policy," are both accurately found and securely filtered (Latenode, 2025).

Effective knowledge management relies on a robust '**Documentation Management Module**' for secure and scalable data ingestion. Consistent with the document ingestion pipeline concepts described by Yonathan (Yonathan et al., 2025), the project adopts a modular architecture where API services utilize a **common contract** for indexing source files. This design abstraction allows the system to remain agnostic to underlying technologies, enabling the seamless swapping of external services (e.g., different OCR engines or embedding providers) and vector database technologies depending on the deployment environment (On-premise or Cloud). Furthermore, security is enforced at the very beginning of the pipeline; access scopes are defined on a partition basis during indexing. This aligns with the metadata-driven resource isolation strategies proposed by Jeong and Lee (Jeong et al., 2025), ensuring that documents are logically isolated and compliant with multi-tenancy requirements from the moment they enter the system.

5.4 Employee Experience and Analytics

LLMs are increasingly used to support employee experience (EX) initiatives and HR analytics by enabling analysis of qualitative feedback at scale. Employee surveys, performance reviews, engagement questionnaires, internal communications, and exit interviews generate substantial volumes of unstructured text that are difficult to analyze using traditional quantitative techniques alone.

A core application is **sentiment and theme analysis**, where LLMs identify recurring topics, concerns, and emotional signals in employee feedback. Compared to classical sentiment analysis, LLMs can capture nuanced opinions, contextual meanings, and mixed sentiments, supporting a more accurate understanding of workforce dynamics.

LLMs also enable **summarization and reporting**, automatically synthesizing large volumes of comments into concise, actionable summaries for HR leaders and managers. State-of-the-art systems increasingly incorporate controllability and traceability mechanisms to reduce misinterpretation risk and improve transparency.

In addition, LLMs can support qualitative HR analytics by contextualizing text feedback alongside structured metrics (e.g., attrition, absenteeism, training outcomes). However, best practice emphasizes aggregation and anonymization to protect privacy and prevent individual-level profiling. Ethical safeguards, transparency to employees, and human judgment remain essential for maintaining trust, particularly when analytics could influence sensitive decisions.

6. Research Gaps and Challenges

While LLMs and immersive technologies such as Extended Reality (XR) offer substantial potential for transforming Human Resources, their practical deployment remains constrained by a set of unresolved technical, organizational, and ethical challenges. In HR contexts, where decisions are high stakes, regulated, and socially sensitive, limitations related to domain adaptation, bias and fairness, explainability, privacy, and system integration become particularly critical. This section synthesizes the key research gaps identified in current state-of-the-art systems and literature, highlighting why existing approaches are insufficient for production-grade HR applications. These challenges directly motivate the design choices and safeguards explored in the Narrate project, which are discussed in the subsequent sections.

6.1 Domain Adaptation

Despite the strong general-purpose capabilities of modern Large Language Models, effective adaptation to the HR domain remains a significant challenge. HR processes rely on organization-specific terminology, internal policies, legal frameworks, and cultural norms that are often underrepresented in large-scale pre-training data. As a result, out-of-the-box models may produce generic or misaligned responses, particularly in high-

stakes workflows such as recruitment, performance evaluation, and employee support, where contextual nuance and policy compliance are essential.

State-of-the-art adaptation strategies, including instruction tuning, parameter-efficient fine-tuning, and Retrieval-Augmented Generation, partially mitigate these limitations but do not fully resolve them. Achieving domain specialization without degrading general language competence remains an open research problem, while the scarcity of high-quality HR datasets limits benchmarking and reproducibility. In addition, HR knowledge is inherently dynamic, as regulations and organizational policies evolve over time, creating a need for controllable and continuously updatable adaptation mechanisms that go beyond one-off fine-tuning and support traceable updates.

Generic LLMs also lack specialized HR terminology, such as labor law references, internal compensation structures, or recruitment-specific jargon. Fine-tuning large models on proprietary HR data is often computationally expensive and risks catastrophic forgetting. Parameter-Efficient Fine-Tuning techniques, such as Low-Rank Adaptation, address this gap by enabling domain adaptation with minimal resource overhead while preserving general capabilities (Hu et al., 2022).

In XR-based deployments, domain adaptation challenges are further amplified by strict real-time constraints. Standalone XR devices must balance the computational demands of automatic speech recognition, LLM inference, and text-to-speech on resource-constrained hardware. Benchmarking studies indicate that smaller models can achieve acceptable on-device generation speeds of approximately 10 to 12 tokens per second, whereas larger models may introduce latency that disrupts conversational flow and immersion (Khan et al., 2025). Maintaining natural interaction therefore requires optimized, streaming-oriented architectures and careful allocation of computation across the multimodal pipeline (Buldu et al., 2025). Together, these constraints highlight that HR domain adaptation is not only a data and modeling challenge but also a deployment concern, particularly in immersive environments where responsiveness directly influences usability and trust.

6.2 Bias and Fairness

While advances in visual realism, driven by technologies such as MetaHuman, have significantly improved the appearance of virtual avatars, achieving behavioral realism remains a substantial challenge. Users may experience discomfort or reduced trust when non-verbal cues such as gestures, gaze, or facial expressions are misaligned with spoken content. Research shows that effective avatars require consistent personality and emotional congruence across verbal and non-verbal channels, which typically demands sophisticated prompt engineering, multimodal coordination, and targeted fine-tuning beyond standard text-centric LLM deployments (Brito et al., 2025; Yamazaki et al., 2023).

In parallel, bias and fairness remain among the most critical and unresolved challenges in applying LLM-based systems to HR contexts. HR decisions directly affect individuals' career trajectories, making even subtle biases in model outputs potentially harmful. Empirical studies demonstrate that LLMs can reproduce and amplify biases present in their training data, including those related to gender, ethnicity, age, and socioeconomic status, which may lead to discriminatory outcomes in recruitment, evaluation, or training recommendations (Mehrabi et al., 2021).

Although state-of-the-art mitigation techniques such as data curation, bias-aware fine-tuning, post-processing interventions, and self-correction workflows have shown partial success, they remain inherently limited and context-dependent. In HR applications, fairness cannot be treated solely as a technical optimization objective, but rather as a socio-technical property shaped by legal requirements, organizational policies, and ethical norms. As a result, defining universal fairness metrics or providing strong model-level guarantees remains difficult.

Consequently, current best practices emphasize bias-aware system design, transparency, and human-in-the-loop oversight rather than full automation. This highlights a persistent gap between model-centric mitigation strategies and the practical demands of real-world HR decision-making, where trust, accountability, and ethical alignment are as critical as predictive performance (Asai et al., 2024).

6.3 Explainability and Interpretability

A significant challenge in applying Large Language Models to HR processes is ensuring explainability and interpretability of AI-driven outputs. HR professionals remain accountable for decisions related to hiring, promotion, or training, which requires a clear understanding of the rationale behind system recommendations. However, LLMs are typically based on large-scale neural architectures that operate as opaque systems, producing predictions or rankings without explicit reasoning paths that are readily interpretable by humans.

This opacity is particularly problematic in HR contexts, where explainability is not only an organizational expectation but often a regulatory requirement. When an AI system ranks candidates or supports evaluative decisions, stakeholders may need to justify outcomes to candidates, management, or regulatory authorities. Without transparent explanations, it becomes difficult to assess fairness, bias, or alignment with organizational policies and ethical standards.

Current approaches to improving transparency, including prompt-based explanations, feature or keyword highlighting, citation-backed generation, and Retrieval-Augmented Generation, provide partial insight into model behavior. These techniques can indicate which documents or CV sections influenced an output and generate natural-language justifications. However, they often fall short of producing explanations that are fully auditable, reproducible, and accessible to non-technical decision-makers.

In high-stakes HR scenarios, such as candidate rejection or performance evaluation, black-box AI outputs are therefore unacceptable. RAG architectures help address this limitation through citation-based generation, in which model responses explicitly reference the source documents used to derive conclusions, enabling human verification of the underlying rationale (Lewis et al., 2020). Nevertheless, a gap remains between current LLM behavior and the level of interpretability required for reliable HR decision support.

Addressing this gap requires hybrid approaches that combine LLM-based reasoning with more interpretable components, such as rule-based logic, structured evaluation criteria, and human-in-the-loop review mechanisms. Incorporating such explainability features is essential for building trust, enabling accountability, and supporting the responsible adoption of AI-assisted decision support in real-world HR settings.

6.4 Privacy and Security

The integration of Large Language Models into XR-based interfaces introduces a distinct class of security and privacy risks that are particularly critical in HR contexts. Unlike traditional text-based systems, XR environments expose additional multimodal input channels, including cameras and microphones, which can be exploited through novel attack vectors. The Evil Vizier study demonstrates how LLM-driven XR assistants may be manipulated via visual prompt injection or audio-based attacks, where malicious instructions are embedded in the avatar's perceptual field or concealed within seemingly benign user input (Zhang et al., 2025). These findings highlight a significant gap in securing multimodal perception channels in real-world HR deployments.

Ensuring robust security in such settings requires safeguards beyond standard LLM defenses. State-of-the-art approaches emphasize multimodal input sanitization, strict content moderation, and clearly defined trust boundaries between perception modules and language reasoning components. Even with these measures, guaranteeing safe operation in open or semi-controlled XR environments remains an open challenge and a key consideration for production-grade systems.

Closely related to security are the stringent privacy and data protection requirements inherent to HR applications. HR systems routinely process highly sensitive personal data, such as employment history, performance evaluations, and benefits-related information. Any use of such data within LLM pipelines introduces risks of data leakage, unauthorized access, and regulatory non-compliance. Compliance with frameworks such as GDPR and local data protection regulations therefore constrains how data is collected, stored, processed, and accessed.

State-of-the-art privacy-preserving practices include data minimization, anonymization, strict access control, and architectural choices such as on-premises or private-cloud deployment. Retrieval-Augmented Generation further reduces privacy risk by limiting direct exposure of raw personal data to model parameters. However, providing end-to-

end privacy guarantees remains challenging, particularly when relying on foundation models with limited transparency. Emerging techniques such as federated learning and privacy-preserving training offer promising directions but are not yet widely adopted in HR-specific systems.

Recent research emphasizes embedding access control directly into RAG architectures rather than relying solely on traditional Role-Based Access Control mechanisms. In these approaches, document-level permissions are enforced during the retrieval phase, ensuring that only authorized content is provided to the LLM as context. Studies report that dynamic, retrieval-time authorization, integrated with identity and access management systems, can effectively prevent unauthorized data exposure, even for highly sensitive information such as health records (Yonathan et al., 2025).

Together, these security and privacy challenges underscore the complexity of deploying LLM-powered XR assistants in HR environments. Addressing them requires a holistic, defense-in-depth strategy that combines technical safeguards, architectural controls, and regulatory compliance, reinforcing the need for cautious and controlled adoption rather than fully autonomous deployment.

6.5 Integration and Interoperability

For an AI assistant to be effective within HR workflows, it must integrate seamlessly with existing enterprise systems such as Human Resource Information Systems and Applicant Tracking Systems. This integration requires not only technical interoperability, including secure access to employee records, job postings, and candidate data, but also alignment with established organizational processes and governance structures. Many AI initiatives fail to move beyond prototypes because they cannot reliably interface with legacy systems or fragmented enterprise data silos.

HR technology ecosystems are inherently heterogeneous, encompassing applicant tracking systems, learning management systems, payroll platforms, and identity and access management solutions. Integrating LLM-based components into such environments introduces significant technical and organizational challenges, including the need for standardized APIs, robust orchestration mechanisms, and fine-grained access control. Without these foundations, LLM-driven services risk remaining isolated tools rather than becoming embedded elements of end-to-end HR workflows.

Equally critical is ensuring that LLMs operate within existing business logic rather than bypassing it. HR decisions require auditability, traceability, and well-defined escalation paths, which are not natively supported by many LLM frameworks. Consequently, AI outputs must be structured, reproducible, and compatible with downstream systems, such as generating standardized candidate summaries or evaluations that can be directly attached to ATS records or reviewed within HR dashboards.

Recent research extends beyond static retrieval-based integration toward agentic workflows, in which LLMs act as orchestrators across enterprise systems. Through function calling capabilities, models can determine whether to retrieve contextual information from a knowledge base or execute API calls to HR systems to access real-time data, such as current leave balances or application status (Dubey et al., 2024). This approach transforms LLMs from passive information providers into functional HR assistants capable of supporting operational tasks.

Addressing these integration challenges is essential for the adoption of LLMs in production HR environments. Solutions such as Narrate must therefore emphasize robust API design, adherence to data exchange standards, and careful output structuring to ensure compatibility with existing tools. Building AI capabilities that complement rather than complicate the HR technology stack is a prerequisite for transitioning from experimental prototypes to sustainable, enterprise-grade HR solutions.

7. Relation to Project Objectives

7.1 Project Overview

The Narrate project aims to design, implement, and validate a set of LLM-enabled solutions that enhance Human Resources processes through intelligent interaction, decision support, and contextual knowledge access. Rather than delivering a single standalone system, Narrate adopts a modular and extensible approach in which multiple AI-driven components address different HR workflows, including recruitment, onboarding, training, knowledge management, and employee support.

Building on recent advances in LLMs, XR technologies, and retrieval-based architectures, the project focuses on embedding language-based intelligence into HR workflows in a manner that is scalable, explainable, and compliant with organizational and regulatory requirements. Narrate explicitly positions LLMs as *augmentative* technologies that support HR professionals, managers, and employees, rather than as fully autonomous decision-makers. Human oversight, transparent system behavior, and controlled use of organizational data are treated as core design principles rather than post hoc safeguards.

Within this framework, components such as conversational assistants for recruitment and knowledge management, XR-based onboarding avatars, analytics modules, and context-aware evaluation tools are developed as HR domain innovations. The Cultural Context Analyzer (CCA) represents one such component, addressing culturally informed interpretation of competencies and qualifications; however, it is not the sole or primary project output. Instead, CCA exemplifies the broader design philosophy of Narrate: combining LLM-based semantic reasoning with structured knowledge sources, domain constraints, and human-in-the-loop control to meet real-world HR requirements.

7.2 Mapping to Use Cases

The state-of-the-art capabilities reviewed in this report directly inform the design and prioritization of Narrate's HR use cases, ensuring that technical choices are grounded in empirical evidence rather than driven by technology push alone. Advances in semantic understanding, conversational alignment, and retrieval-based grounding enable HR solutions that move beyond traditional rule-based or keyword-driven systems toward more adaptive, context-aware interactions.

In the context of intelligent recruitment and interviewing, Narrate leverages avatar-based candidate interaction to support structured yet natural interview experiences. Research conducted on the Milo platform demonstrates that complex social interactions such as interviews benefit from human-in-the-loop architectures, where AI-driven agents operate under human oversight to ensure ethical compliance, contextual appropriateness, and fairness (Shoa & Friedman, 2025). This approach aligns with Narrate's goal of augmenting, not replacing, human decision-making in high-stakes HR processes.

For immersive onboarding and training, Narrate's use of XR-based procedural guidance is supported by findings from the XaiR and agentAR studies. These works show that multimodal LLMs, when combined with AR overlays, can effectively guide users through physical and procedural tasks by providing contextual, step-by-step assistance in real time (Srinidhi et al., 2024; Zhu et al., 2025). Such capabilities are particularly relevant for onboarding scenarios, where employees must quickly acquire operational knowledge while interacting with real-world environments.

In the domain of HR knowledge management, Narrate adopts a RAG pipeline for accessing internal policies, guidelines, and procedural documentation. This design is directly informed by the architecture proposed by Tomkou et al. (2025), which emphasizes semantic chunking to enable high-fidelity retrieval of technical information in industrial XR settings. Applying this approach to HR content ensures that generated responses remain accurate, traceable, and grounded in authoritative organizational sources.

Across these use cases, techniques such as instruction tuning and RAG form the foundation for recruitment and selection tasks that require precise interpretation of job requirements and candidate profiles. Likewise, conversational LLMs aligned through RLHF support employee-facing onboarding and HR support scenarios by delivering consistent, policy-aligned guidance. For employee experience analytics, LLM-based summarization and thematic analysis enable the extraction of actionable insights from qualitative feedback while respecting privacy, fairness, and ethical constraints.

By systematically mapping state-of-the-art LLM and XR capabilities to concrete HR workflows, Narrate ensures that AI adoption is driven by demonstrable operational value, regulatory awareness, and user trust, establishing a clear bridge between advanced research and practical, production-ready HR systems.

Addressing the Precision Gap via Hybrid Search

Standard semantic search often fails to retrieve specific HR codes or exact procedure names (zero-shot scenarios). To address this identified gap, the project plans to utilize a Hybrid Search architecture. This significantly improves recall compared to single-method approaches by enabling the system to process both conceptual queries (e.g., ‘Maternity leave entitlements’) and exact keyword searches (e.g., ‘Form 4A’).

Solving the Unstructured Data Challenge with Multimodal Ingestion or OCR

Traditional HR knowledge bases struggle to cope with PDFs containing organizational charts, salary tables and scanned policies. The project addresses this gap by adopting Multimodal Embedding (VLM) and OCR techniques. By processing documents visually rather than just textually, the system preserves layout and context, enabling the retrieval of information embedded in complex visual formats which standard text-based RAG systems miss.

Mitigating Data Leakage with Retrieval-Time Security

A critical security vulnerability may arise in HR environments where sensitive data (e.g., managers' salaries, disciplinary records) is stored alongside general policies. To address this weakness, the project adopts a Role-Based Access Control (RBAC) system. This ensures that users can only retrieve context they are explicitly authorized to view, effectively preventing unauthorized data leakage during semantic search.

7.3 Expected Innovations

Narrate is expected to deliver a set of interrelated innovations at both the technical and application levels, structured around modular, interoperable components rather than a single monolithic solution. Each component addresses a specific gap identified in current HR-focused LLM deployments, while collectively contributing to a scalable, trustworthy, and enterprise-ready HR platform.

CV Analysis and Job Description Competency.

Narrate introduces an LLM-assisted CV analysis module that systematically evaluates candidate qualifications against job-specific competency requirements through requirement-level semantic reasoning. CVs in PDF or DOCX format are automatically parsed to extract technical skills, soft skills, experience indicators, and domain-relevant keywords, which are then matched against each job requirement individually rather than via coarse similarity scoring. The module produces an interpretable competency alignment score alongside explanatory outputs, including identified skill gaps and a structured summary of salient candidate strengths. By combining multilingual (Turkish and English) semantic understanding with transparent evaluation logic, this component delivers actionable, auditable insights that support recruiter decision-making while aligning with fairness and regulatory expectations.

Real-time Emotion Recognition for Interviews.

Narrate advances interview intelligence through real-time emotion recognition applied during live interview sessions, enabling adaptive and context-aware candidate interactions. Detected emotional signals such as confidence, hesitation, stress, or engagement are continuously fed into the recruitment chatbot to dynamically adjust question difficulty, pacing, and interview flow. This adaptive mechanism allows the system to modulate the intensity of the interview while extracting complementary insights into candidate responses beyond verbal content alone. The innovation lies in integrating real-time affective signals as decision-support inputs rather than evaluative judgments, thereby enriching interview context while preserving human oversight and ethical constraints.

LLM-Enabled Recruitment and Interview Support Components.

Narrate introduces an LLM-based, agentic recruitment chatbot designed to manage and support interview processes across multiple structured phases aimed at assessing both technical and soft skills. The chatbot orchestrates the interview flow by integrating inputs from upstream components, including job-specific competency requirements and extracted candidate skills from the CV analysis module, as well as real-time affective signals provided by the emotion recognition component. Based on this combined contextual understanding, the system dynamically selects and adapts interview questions, including CV-driven clarifications, position-specific inquiries, and behavioral prompts, while adjusting pacing and difficulty in response to candidate engagement and emotional state.

Following the assessment phase, the chatbot generates a structured interview report and competency-based scoring output intended for HR recruitment professionals. These outputs are designed to be explanatory and auditable, providing insight into how observed responses, competencies, and interaction dynamics contributed to the evaluation. The innovation lies in the use of an agentic, multi-input LLM architecture that coordinates semantic reasoning and adaptive interaction as decision support rather than autonomous judgment, thereby ensuring human oversight, transparency, and alignment with fairness and regulatory requirements in high-stakes recruitment contexts.

Cultural Context Analyzer (CCA).

The Cultural Context Analyzer represents a dedicated project-level component focused on culturally informed interpretation of competencies, qualifications, and communication styles. By integrating LLM-based semantic analysis with structured skills frameworks (e.g., ESCO) and cultural ontologies, the CCA accounts for variations in how experience and competencies are expressed across cultural contexts. Rather than serving as a standalone product, the CCA exemplifies Narrate’s broader approach to embedding contextual awareness into HR decision-support tools. Its innovation lies in enabling more equitable and nuanced evaluation while preserving transparency and human oversight.

XR–LLM Integrated Onboarding and Training Assistants.

Narrate advances immersive onboarding through the integration of LLMs with XR-based interfaces, enabling intelligent digital assistants that provide contextual, procedural guidance in real time. These assistants combine conversational interaction with spatial awareness and visual overlays to support hands-on learning and task execution. Drawing on recent research in XR latency optimization and multimodal interaction, Narrate explores modular architectures that balance responsiveness, accuracy, and device constraints. The innovation lies in transforming static onboarding materials into adaptive, interactive experiences while addressing latency, usability, and trust challenges inherent in immersive environments.

HR Knowledge Management and Support Agents.

Another key innovation is the deployment of retrieval-augmented LLM agents for HR knowledge management and employee support. These agents provide natural language access to internal policies, procedures, and guidelines, grounding responses in verified organizational documents to ensure accuracy and traceability. This approach enables HR knowledge to be treated not merely as static documentation, but as a queryable, context-aware, and continuously updatable knowledge structure. By adopting semantic chunking and controlled retrieval strategies, Narrate minimizes hallucination risk and supports auditability and critical requirements in HR contexts. This component demonstrates how LLMs can be safely integrated into enterprise knowledge workflows as reliable, low-risk support tools. Linking generated responses to their underlying source documents provides significant potential in terms of **transparency and explainability**, while also supporting retrospective review and traceability needs in HR processes.

The Narrate architecture is designed to support the presentation of the same information at different levels of detail depending on user roles and usage contexts. This allows employees, managers, and HR professionals to benefit from the same organizational knowledge base in a manner aligned with their respective needs, contributing to improved clarity in complex HR processes.

In addition, this structure supports the semi-automated handling of recurring HR-related queries, thereby contributing to the reduction of operational workload for HR teams and creating capacity for more strategic and value-added activities.

Employee Experience and Analytics Components.

Narrate also explores LLM-based components for analyzing qualitative employee feedback and supporting HR analytics. These components apply advanced summarization, thematic analysis, and sentiment interpretation to large volumes of unstructured text, such as survey responses or open-ended feedback. This approach enables qualitative employee feedback to be addressed in a more systematic and consistent manner, rather than relying on individual interpretation or manual evaluation processes. In particular, the assessment of large volumes of open-ended feedback from

multiple sources within a common analytical framework supports the generation of comparable and holistic insights across the organization. Innovation is achieved through aggregation-focused design, anonymization, and transparency mechanisms that protect employee privacy while enabling evidence-based insights. Rather than replacing human judgment, these tools highlight emerging trends and potential issues to inform HR decision-making responsibly. The aggregation-focused analytical structure supports the identification and monitoring of emerging patterns and recurring themes over time. By shifting the focus from individual feedback items to collective trends, this approach allows structural issues and potential risk areas affecting employee experience to become visible at earlier stages.

Anonymization and privacy-oriented design principles place ethical use and data protection at the center of employee feedback analysis. This perspective supports employee trust in feedback mechanisms while contributing to the long-term sustainability of analytics-driven employee experience initiatives. The developed analytical components aim not to automate final decision-making, but to provide evidence-based decision-support inputs for HR teams. This enables analytical outputs to be interpreted in conjunction with human judgment, reinforcing a responsible and context-aware use of analytics.

The resulting thematic and summarized outputs may be leveraged as inputs for organizational learning and continuous improvement initiatives in the area of employee experience. In this way, employee feedback moves beyond one-off measurement exercises and exhibits potential to evolve into a continuously monitored and informant knowledge source over time.

Cross-Cutting Technical and Governance Innovations.

Across all components, Narrate introduces cross-cutting innovations in system governance and deployment. These include retrieval-augmented and policy-aware architectures, parameter-efficient adaptation strategies, explainability-oriented design, and robust integration mechanisms with existing HR systems. Special attention is given to security and privacy in multimodal and XR settings, incorporating input sanitization, defensive prompting, and controlled data access. Together, these design choices reflect a defense-in-depth approach to trustworthy AI deployment in HR.

Collectively, these component-level innovations position Narrate as a platform-oriented contribution that demonstrates how multiple LLM-enabled solutions can be responsibly orchestrated across HR workflows. By addressing technical, ethical, and organizational challenges in an integrated manner, Narrate bridges the gap between state-of-the-art AI research and production-grade HR systems.

8. Conclusions

8.1 Summary of Findings

This report has reviewed the state of the art in Large Language Models with a focus on their relevance to Human Resources applications. Recent advances in transformer-based architectures, large-scale pre-training, instruction tuning, and alignment techniques have positioned LLMs as powerful general-purpose language technologies capable of supporting complex, knowledge-intensive workflows.

The analysis shows that LLMs are particularly well-suited for HR processes that rely on unstructured textual data and conversational interaction, such as recruitment, onboarding, knowledge management, employee training, and employee experience analysis. Techniques such as retrieval-augmented generation and parameter-efficient fine-tuning enable these models to be adapted to organizational contexts while improving accuracy, controllability, and compliance.

At the same time, the report highlights persistent challenges related to bias, explainability, privacy, and system integration. These challenges underscore the importance of responsible AI practices and human-in-the-loop system design, especially in HR contexts where automated outputs may have significant social and legal implications.

8.2 Readiness Level of LLMs for HR

From a technological perspective, LLMs have reached a level of maturity that supports their deployment as decision-support and interaction tools in HR environments. Applications such as HR chatbots, knowledge assistants, and document analysis systems are already viable and, in some cases, operationally deployed in enterprise settings.

However, full automation of HR decision-making remains neither technically robust nor ethically acceptable. The current readiness level favors hybrid systems that combine LLM capabilities with structured workflows, organizational knowledge bases, and human oversight. Models must be carefully adapted, evaluated, and governed to ensure fairness, transparency, and regulatory compliance.

Overall, LLMs can be considered ready for selective and controlled adoption in HR, with clear boundaries on their role and responsibilities.

8.3 Implications for the Project

For the Narrate project, these findings confirm the feasibility and relevance of LLM-based approaches to enhancing HR processes. The state-of-the-art evidence supports Narrate’s emphasis on retrieval-augmented architectures, responsible AI principles, and integration with existing HR systems.

By focusing on augmentation rather than automation, and by embedding governance and explainability into system design, Narrate is well positioned to address current

research gaps and deliver practical, trustworthy HR solutions. The project's outcomes are expected to contribute both to applied innovation and to broader understanding of how LLMs can be responsibly deployed in human-centric enterprise domains.

9. References

Asai, A., Wu, Z., Wang, Y., Sil, A., & Hannaneh, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. arXiv preprint arXiv:2310.11511.

Amazon Web Services. (2023, May 2). Integrate sparse and dense vectors to enhance knowledge retrieval in RAG using Amazon OpenSearch Service. AWS Big Data Blog. https://aws.amazon.com/tr/blogs/big-data/integrate-sparse-and-dense-vectors-to-enhance-knowledge-retrieval-in-rag-using-amazon-opensearch-service/

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. https://doi.org/10.1145/3442188.3445922

Boumans, R., Cramer, L., van de Poll, S., & Vermeulen, H. (2025). A feasibility study on usability and trust among population groups of a medical avatar supported by large language models with retrieval augmented generation. Behavioural Science Institute, Radboud University.

Brito, I. A., Färber, F. B., et al. (2025). Integrating personality into digital humans: A review of LLM-driven approaches for virtual reality. arXiv (Computer Science – Human-Computer Interaction).

Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Buldu, K. B., Özdel, S., et al. (2025). CUIfy the XR: An open-source package to embed LLM-powered conversational agents in XR. 2025 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR).

Chen, B., Tackman, J., Setälä, M., Poranen, T., & Zhang, Z. (2025). Integrating access control with retrieval-augmented generation: A proof of concept for managing sensitive patient profiles. Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing (SAC '25), 915–917. https://dbs-research.github.io/pdf/2025_sac.pdf

Chowdhery, A., Narang, S., Devlin, J., et al. (2022). PaLM: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311.

Cordioli, L., Piro, L., Valoriani, M., & Matera, M. (2025). Exploring LLM-driven interaction for knowledge retrieval in extended reality. CHItaly 2025 (Proceedings of the 16th Biannual Conference of the Italian SIGCHI Chapter).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dubey, A., Jauhri, A., Pandey, A., Keshwam, A., … & Meta AI. (2024). The Llama 3 herd of models. arXiv preprint arXiv:2407.21783.

Faysse, M., Sibille, H., Wu, T., Omrani, B., Viaud, G., … & Hudelot, C. (2024). ColPali: Efficient document retrieval with vision language models. arXiv preprint arXiv:2407.01449.

Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research, 23(120), 1–39.

Frantar, E., Stock, P., & Alistarh, D. (2022). GPTQ: Accurate post-training quantization for generative pre-trained transformers. arXiv preprint arXiv:2210.17323.

Gianni, A. M., Nikolakis, N., & Antoniadis, N. (2025). An LLM-based learning framework for adaptive feedback mechanisms in gamified XR. Computers & Education: X Reality.

Hendrycks, D., Burns, C., Basart, S., et al. (2021). Measuring massive multitask language understanding. Proceedings of ICLR 2021.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hoffmann, J., Borgeaud, S., Mensch, A., et al. (2022). Training compute-optimal large language models. arXiv preprint arXiv:2203.15556.

Hu, E. J., Shen, Y., Wallis, P., et al. (2022). LoRA: Low-rank adaptation of large language models. Proceedings of ICLR 2022.

Hui, B., Yang, Z., Yu, L., … & Qwen Team. (2025). Qwen2.5-Coder: The code evolution. arXiv preprint arXiv:2412.15115.

Jeong, J., & Lee, S.-G. (2025). Permission-aware RAG: Identity and access management (IAM)-based access filtering in multi-resource environments. IEEE Access.

Jiang, A., Sablayrolles, A., Mensch, A., et al. (2023). Mixtral of experts. arXiv preprint arXiv:2401.04088.

Jurafsky, D., & Martin, J. H. (2009). Speech and language processing (2nd ed.). Pearson.

Kaplan, J., McCandlish, S., Henighan, T., et al. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

Khan, D., Liu, X., Mena, O., Jia, D., Kouyoumdjian, A., & Viola, I. (2025). AIvaluateXR: An evaluation framework for on-device AI in XR with benchmarking results. arXiv (Preprint).

Latenode. (2025). Best vector databases for RAG: Complete 2025 comparison guide. Latenode Blog. https://latenode.com/blog/ai-frameworks-technical-infrastructure/vector-databases-embeddings/best-vector-databases-for-rag-complete-2025-comparison-guide

Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

Liang, P., Bommasani, R., Lee, T., et al. (2022). Holistic evaluation of language models. arXiv preprint arXiv:2211.09110.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35. https://doi.org/10.1145/3457607

Microsoft. (2023, September 23). Azure AI Search: Outperforming vector search with hybrid retrieval and reranking. Microsoft Tech Community Blog. https://techcommunity.microsoft.com/blog/azure-ai-foundry-blog/azure-ai-search-outperforming-vector-search-with-hybrid-retrieval-and-reranking/3929167

Mistral AI. (2024, July). Mistral NeMo: Our new best-in-class small model. Mistral AI News. https://mistral.ai/news/mistral-nemo

Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.

Qdrant. (2024). Vector database benchmarks. Qdrant Technical Reports. https://qdrant.tech/benchmarks/

Radford, A., Wu, J., Child, R., et al. (2019). Language models are unsupervised multitask learners. OpenAI Technical Report.

Raffel, C., Shazeer, N., Roberts, A., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1–67.

Rafailov, R., Sharma, A., Mitchell, E., et al. (2023). Direct preference optimization: Your language model is secretly a reward model. arXiv preprint arXiv:2305.18290.

Shoa, A., & Friedman, D. (2025). Milo: An LLM-based virtual human open-source platform for extended reality. Frontiers in Virtual Reality.

Song, Y., & Xiong, W. (2025). Large language model-driven 3D hyper-realistic interactive intelligent digital human system. Sensors.

Srinidhi, S., Lu, E., & Rowe, A. (2024). XaiR: An XR platform that integrates large language models with the physical world. IEEE ISMAR.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. NeurIPS Datasets and Benchmarks Track.

Tomkou, D., Fatouros, G., Andreou, A., et al. (2025). Bridging industrial expertise and XR with LLM-powered conversational agents. arXiv.

Touvron, H., Lavril, T., Izacard, G., et al. (2023). LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

Wei, J., Bosma, M., Zhao, V. Y., et al. (2021). Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652.

Yamazaki, T., Mizumoto, T., Yoshikawa, K., Ohagi, M., Kawamoto, T., & Sato, T. (2023). An open-domain avatar chatbot by exploiting a large language model. SIGDIAL.

Yonathan, M. (2025). Access-controlled semantic search: Implementing role-based filtering in vector databases for enterprise document management. Research Square.

Zhang, Y., Shayegani, E., Huang, Z., Chen, J., Chen, S., & Abu-Ghazaleh, N. (2025). Evil Vizier: Vulnerabilities of LLM-integrated XR systems. arXiv.

Zheng, L., Chiang, W.-L., Sheng, Y., et al. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv preprint arXiv:2306.05685.

Zhu, C., Hsia, S.-K., Hu, X., Liu, Z., Shi, J., & Ramani, K. (2025). agentAR: Creating augmented reality applications with tool-augmented LLM-based autonomous agents. UIST ’25.

10. Appendix

10.1 Overview of Key LLM Benchmarks

The following table summarizes widely used benchmarks for evaluating Large Language Models, highlighting their focus areas and relevance to HR-oriented applications.

Table 2: Widely used benchmarks for LLMs and relevance to HR domain

Benchmark	Primary Focus	Evaluation Type	Relevance to HR Applications
MMLU (Hendrycks et al., 2021)	General knowledge and reasoning across multiple domains	Multiple-choice QA	Useful for assessing general reasoning and knowledge breadth, but limited for conversational HR tasks
HELM (Liang et al., 2022)	Holistic model evaluation (accuracy, bias, robustness, toxicity, efficiency)	Multi-metric framework	Highly relevant due to its emphasis on fairness, robustness, and transparency
MT-Bench (Zheng et al., 2023)	Instruction following and open-ended dialogue	LLM-as-a-judge + human preference	Relevant for evaluating conversational HR assistants and chatbots
Chatbot Arena (LMSYS, 2023)	Real-world user preference comparison	Crowdsourced pairwise ranking	Provides insights into perceived usefulness and quality in interactive settings
TruthfulQA	Hallucination and factual correctness	Open-ended QA	Relevant for HR policy and compliance-related use cases

Observation:
No existing benchmark fully captures HR-specific risks such as hiring bias, policy compliance, or privacy handling, underscoring the need for domain-aware evaluation protocols in HR deployments.

10.2 Mapping of LLM Capabilities to HR Use Cases

The following matrix illustrates how state-of-the-art LLM capabilities map to key HR processes addressed in this report and within the Narrate project.

Table 3: How state-of-the-art LLM capabilities map to key HR processes within the Narrate project

LLM Capability	Recruitment &	Onboarding &	Knowledge	Employee

	Selection	Training	Management	Experience & Analytics
Semantic text understanding	CV screening, role–candidate matching	Role-specific learning content	Policy interpretation	Feedback analysis
Conversational interaction	Candidate chatbots	Onboarding assistants	HR helpdesk support	Employee pulse surveys
Instruction following	Interview question generation	Guided learning workflows	Structured HR guidance	Survey summarization
Retrieval-Augmented Generation (RAG)	Job description grounding	Training material retrieval	Policy and FAQ access	Contextual insight generation
Summarization	Interview notes	Training progress summaries	Policy updates	Engagement and sentiment reports
Bias-aware prompting & controls	Fairer screening support	Inclusive training content	Neutral policy responses	Aggregated, anonymized insights

Key insight:
LLMs deliver the greatest value in HR when used as **augmentative systems**, supporting human decision-makers rather than replacing them, particularly in high-risk processes such as recruitment and evaluation.

10.3 Identified Gaps for Future HR-Specific LLM Evaluation

Based on the state-of-the-art review, the following gaps remain insufficiently addressed by current benchmarks and tooling:

- Lack of **standardized HR-specific evaluation datasets**
- Limited measurement of **fairness under realistic recruitment scenarios**
- Insufficient testing of **policy compliance and explainability**
- Weak support for **privacy-aware evaluation**, particularly with personal data
- Absence of longitudinal evaluation for **organizational impact over time**

These gaps motivate the need for **custom evaluation protocols and pilot studies**, as envisioned within the Narrate project.