A Benchmark and Field Study on the Use of LLMs and RAG for Harmonizing Job Descriptions in the Austrian Federal Administration*

Nicolas Ferranti[1,*,†], Andreas Mild[1,†] and Stefan Sobernig[1,†]

[1]Vienna University of Economics and Business, Welthandelsplatz 1, 1020, Vienna, Austria

Abstract

Job descriptions play a central role in public-sector personnel management, job grading, and organizational development. However, in many administrations they remain fragmented, manually maintained, and difficult to compare across organizational units, limiting transparency and hindering digital transformation. In the Austrian federal public administration, these challenges are compounded by heterogeneous regulatory frameworks and the lack of structured, interoperable job description repositories. This paper explores the use of Large Language Models (LLMs) combined with Retrieval-Augmented Generation (RAG) to support the semi-automatic generation and harmonization of structured job description components. We propose a RAG-based pipeline that generates goals, tasks, activities, and requirements grounded in existing administrative documents and expert-authored descriptions. The approach is evaluated through two studies. First, we perform an automatic summarization and similarity analysis comparing multiple embedding-based, lexical, and n-gram-based methods across job description components. Second, we conduct an expert-based evaluation assessing linguistic quality, correctness, and usefulness of the generated outputs. Results indicate that no single similarity measure performs best across all content types. SBERT-based embeddings achieve the most consistent performance, reaching 96% rank-based classification accuracy for task descriptions, while expert assessments confirm high linguistic quality but exhibit limitations for more abstract text elements.

Keywords

Job Description, Large Language Model, Retrieval-Augmented Generation, Text Summarization, Text Similarity, Benchmark Study, Field Study

1. Introduction

Technological development has reshaped knowledge-intensive work, including the progressive digitalization of public administration and its interactions with citizens and stakeholders [1]. However, many internal workflows remain largely manual, leading to inefficiencies, inconsistencies across administrative units, and reliance on individual expertise, which limits their potential to support systematic decision-making [2].

Job descriptions (JDs) are a widely used instrument in human resource management in organizations based on the division of labor. While their exact structure may vary across institutions, most JDs specify (i) the classification of a position within the organizational structure and (ii) the tasks, objectives, and responsibilities associated with that position [3]. In the Austrian federal public administration, JDs play a particularly important role because they serve as a formal basis for personnel appraisal and remuneration decisions. In principle, each position in the general administrative service must be defined through a JD. These descriptions are used by the Federal Chancellery to evaluate the position and classify it into a specific usage group (*Verwendungsgruppe*) and function group (*Funktionsgruppe*), which determine key remuneration components and career levels [4].

To ensure fair and consistent evaluation across ministries and administrative units, JDs must support

3rd Workshop on Evaluation of Language Models in Knowledge Engineering (ELMKE) co-located with ESWC, May 10–14, 2026, Dubrovnik, Croatia

*Corresponding author.

†These authors contributed equally.

nicolas.ferranti@wu.ac.at (N. Ferranti); andreas.mild@wu.ac.at (A. Mild); stefan.sobernig@wu.ac.at (S. Sobernig)

0000-0002-5574-1987 (N. Ferranti); 0009-0003-5020-5362 (A. Mild); 0009-0002-5018-7961 (S. Sobernig)

both horizontal comparability and vertical alignment. Horizontal comparability ensures that similar roles across departments can be evaluated consistently, supporting fairness and transparency in public-sector personnel management. Vertical alignment connects individual responsibilities and goals to higher-level strategic objectives and legal mandates within the administrative hierarchy [5].

Despite their importance, current JD management practices in many administrations remain fragmented and largely analogue. JDs are typically created by job holders, reviewed by supervisors, and subsequently evaluated by central authorities. This multi-stage process often results in heterogeneous documents that vary in structure, terminology, and level of detail [6]. In practice, JDs are frequently stored as isolated Word or PDF files without standardized metadata, centralized repositories, or systematic version control. This fragmentation limits interoperability, complicates cross-organizational comparison, and constrains the integration of job descriptions into digital HR systems.

Similar challenges have been observed in broader labor market contexts, where heterogeneous job descriptions and inconsistent occupational data complicate large-scale analysis and workforce planning [7]. Initiatives such as the European Skills, Competences, Qualifications and Occupations (ESCO) framework aim to address these issues by providing standardized taxonomies for occupations and skills [8]. However, the integration of such frameworks into public-sector administrative workflows remains largely manual and requires substantial expert effort. Recent advances in Large Language Models (LLMs) have demonstrated strong capabilities in text summarization, information extraction, and structured text generation [9, 10, 11]. These capabilities suggest potential support for knowledge-intensive administrative tasks such as consolidating task descriptions, abstracting goals, or standardizing qualification requirements across heterogeneous documents. However, LLMs are not inherently grounded in legal, regulatory, or organizational frameworks and may produce plausible but inaccurate or incomplete outputs. Such limitations are particularly critical in regulated domains such as public administration, where institutional accountability is essential [12].

To mitigate these risks, recent research has proposed Retrieval-Augmented Generation (RAG), which constrains LLM outputs by grounding generation in authoritative domain-specific sources [13]. By retrieving relevant documents or document fragments from curated corpora, RAG enables language models to generate outputs that remain aligned with existing institutional knowledge while reducing the risk of hallucination.

Evaluating the effectiveness of such systems requires criteria beyond surface-level textual fluency. In structured domains such as public-sector job descriptions, evaluation must consider semantic alignment with source documents, structural consistency of generated content, and practical usefulness for administrative workflows. Automatic similarity metrics can provide quantitative indicators of semantic correspondence, but they cannot fully capture contextual appropriateness or institutional adequacy. Therefore, expert-based evaluation remains essential to assess linguistic correctness, factual accuracy, and practical usability in real-world settings.

Within this context, this paper investigates how RAG can support the semi-automatic generation and harmonization of structured job description components in the Austrian federal administration. The proposed approach combines embedding-based retrieval, controlled LLM prompting, and expert validation to assist administrative experts in drafting consistent and comparable job descriptions.

The research project is guided by the following research questions:

RQ_1 *Do the generated and reference job descriptions match? To this end, we investigate whether:*
- RQ_{1a} Which text-similarity construct provides a useful proxy for semantic matching between generated and reference job descriptions?
- RQ_{1b} Is the usefulness of similarity constructs related to the large-language model size (i.e., number of parameters)?

RQ_2 *How do domain experts judge the quality (language, correctness, usefulness) of the generated job descriptions?*

The main contributions of this paper are fourfold. First, we propose a RAG-based pipeline architecture tailored to the generation and harmonization of structured public-sector job descriptions. Second,

we evaluate multiple similarity constructs for assessing semantic alignment between generated and reference descriptions. Third, we provide an empirical assessment of the feasibility of integrating LLM-assisted workflows into public-sector HR processes through an expert-based evaluation. Fourth, we make all permissible materials available on GitHub.[1]

2. Background

Job descriptions (JDs) are central artifacts in public-sector personnel management, job grading, and organizational development. Beyond their administrative role, they encode information about organizational roles, responsibilities, goals, and qualification requirements [3]. From a knowledge engineering perspective, JDs can therefore be viewed as semi-structured knowledge objects that formalize institutional expectations across organizational units. In public administration, they must ensure both horizontal comparability across departments and vertical alignment with strategic objectives and legal mandates, making them inherently multidimensional representations of tasks, goals, and regulatory constraints [5].

JDs are central to the formal job evaluation process in the Austrian federal administration. Positions are assessed according to constructs such as knowledge requirements, intellectual effort, and responsibility levels, which are further divided into evaluation dimensions (e.g., specialist knowledge or managerial responsibility) [14]. Based on standardized descriptions and point scales, positions are classified and assigned to specific usage and function groups within the civil service remuneration system.

The creation and maintenance of JDs involves multiple actors, including job holders, supervisors, and central administrative units. This multi-stage process often leads to heterogeneous and loosely structured documents with varying formulations and limited metadata [6]. As a result, interoperability and systematic analysis remain difficult. From a knowledge engineering perspective, this fragmentation reflects the lack of a unified, machine-interpretable representation, a challenge also highlighted by the Austrian Court of Audit [4].

Large Language Models (LLMs) have demonstrated strong performance in several natural-language processing tasks including generation, summarization, and extraction [9, 10, 11], making them promising tools for knowledge-intensive domains [15]. In the context of job descriptions, LLMs can support the consolidation of task lists, abstraction of goals, and standardization of qualification requirements.

From a knowledge engineering perspective, LLMs act as generative interfaces over latent semantic representations learned from large corpora [16]. They can approximate semantic similarity and produce coherent reformulations aligned with domain conventions. However, they also exhibit well-documented limitations, including hallucinations, omission of constraints, and insufficient institutional grounding [12]. These risks are particularly critical in regulated domains such as public administration, where legal compliance, traceability, and procedural correctness are essential.

Therefore, purely generative approaches are insufficient for accountable knowledge production. LLMs must be embedded in controlled and human-moderated workflows that constrain generation and enable systematic validation, motivating the use of retrieval-based grounding mechanisms and human-in-the-loop evaluation.

Retrieval-Augmented Generation (RAG) integrates generative language models with a retrieval component that selects relevant documents from a curated corpus [13]. Rather than relying solely on internal model representations, generation is conditioned on externally retrieved, domain-specific sources. This grounding reduces hallucination risk and improves traceability by linking outputs to explicit evidence [17].

From a knowledge engineering perspective, RAG represents a hybrid architecture combining symbolic retrieval mechanisms (e.g., indexed corpora and embedding-based similarity search) with subsymbolic generative models. Retrieval provides contextual grounding, while generation synthesizes structured

[1] https://github.com/nicolasferranti/rag-llm-generated-job-descriptions, last accessed: March 17th, 2026

outputs aligned with the target representation. In the case of JDs, RAG enables the reuse of expert-authored documents as references when generating harmonized goals, tasks, or requirements.

Text summarization is a core task in Natural Language Processing (NLP) that aims to condense longer texts into shorter representations while preserving their essential meaning [18]. Approaches are distinguished between *extractive* summarization, which selects relevant segments from the source text, and *abstractive* summarization, which generates new formulations that capture the underlying semantic content [19]. Recent LLM-based methods have improved abstractive summarization by leveraging large-scale pretrained language representations [9]. In the context of job descriptions, summarization techniques can support the consolidation of heterogeneous task lists and responsibility statements into more concise and comparable formulations. Consequently, summarization provides a suitable generative mechanism for harmonizing job description components across heterogeneous documents.

Text similarity measures the degree to which two textual representations convey related semantic content. Traditional approaches rely on lexical overlap metrics such as cosine similarity over TF–IDF vectors or n-gram-based measures [20]. More recent methods employ neural embeddings that map texts into high-dimensional semantic vector spaces, enabling similarity comparisons based on contextual meaning rather than surface-level word overlap [21]. Such embedding-based similarity constructs have become widely used for evaluating generated text, as they provide quantitative indicators of semantic correspondence between generated and reference texts. In the context of job descriptions, similarity measures can therefore serve as proxies for assessing how closely LLM-generated descriptions align with expert-authored reference descriptions.

3. A RAG Pipeline for Job-Description Summarization

To support the semi-automatic generation of job-description summaries to harmonize public-sector job descriptions and to support a human-participant evaluation, we designed, implemented and deployed a functional experimental prototype offering a Jupyter-based Web UI on top of the RAG pipeline. Figure 1 illustrates the main components of the system's architecture and user-driven workflow implemented by the pipeline. The experimental prototype integrates structured data storage, embedding-based retrieval, LLM-based generation, and human-in-the-loop moderation, and data collection from human participants. The prototype realizes a hybrid knowledge-engineering architecture that combines symbolic data structuring and retrieval with subsymbolic generation. By integrating LLM-based text generation in a moderated, human-controlled workflows and expert oversight, the pipeline supports validating and consolidating job description while maintaining institutional requirements.

Test data A corpus of 1,337 job descriptions (JDs) from the Austrian federal public administration was available from our project partner. This corpus is the result of a prior data-processing, tidying, and data-enrichment activity. Processing involved sourcing the JD text segments from heterogeneous data files (e.g., Microsoft Word Binary and Microsoft Word Open XML files) into database records. In addition, processing included cleaning and re-composing text segments into semantically coherent text units: goals, tasks, activities, and qualification requirements. Detecting and extracting tabular and itemized textual data was required, among others. Data enrichment required to extract or to retrieve from additional data sources and, manually, map additional administrative metadata (descriptions of roles, functions, organizational units etc.) to the tidied job descriptions. The tidied dataset was stored in a PostgreSQL database serving as the retriever component. A predefined database schema reflecting the JD structure and entities for section, department, function group, and other organizational attributes was used. The data-processing pipeline as well as the database schema can be reused to curate additional datasets, including Austrian public-sector job announcements.

Infrastructure The prototype and its components were deployed in a distributed setup, including two dedicated GPU servers:

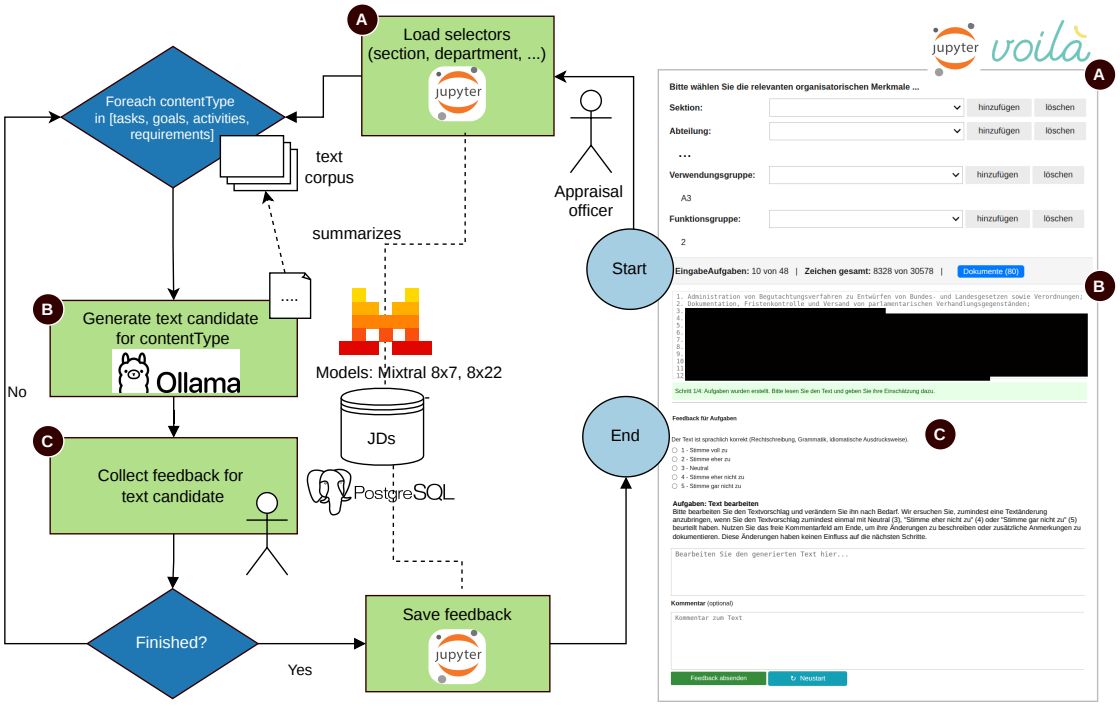

Figure 1: Overview of the user-driven workflow to generate and to assess job descriptions, based on a Web user interface (Jupyter, voilà) on top of a RAG pipeline (Mixtral models, Ollama, PostgreSQL).

	Jupyter, Ollama (Mixtral 8x7B)	Ollama (Mixtral 8x22B)
OS	Ubuntu 22.04.5 LTS	Ubuntu 24.04.3 LTS
GPU	4x NVIDIA RTX A5000 (paired), 24GB	1x NVIDIA H100L-94C, 96GB
CPU	AMD EPYC 9334	AMD EPYC-Milan
RAM	90GB	126GB

Retriever and Summary Generator The RAG pipeline is organized into three main stages: parameter selection, retrieval, and summary generation. In an interactive UI task (Figure 1), appraisal officers select contextual parameters such as organizational unit, department, or function group. These parameters are used to select relevant job description components from the database. Text units of the selected job descriptions serve as a grounding context for summary generation.

The models are prompted using a fixed, predefined meta-prompt [22] specifying the output structure, syntax and style generically (see Appendix C). The meta-prompt specifies formatting requirements (e.g., list structure with delimiter constraints), maximum length, language restrictions (German only), domain-specific terminology, and stylistic limitations such as avoiding filler phrases and redundancy. These constraints aim to ensure that generated components remain concise, standardized, and aligned with administrative drafting conventions. Meta-prompting was chosen over alternatives (e.g., few-shot prompting) to minimise the prompt's token consumption itself and to enable a comparison of different models, avoiding bias through output examples.

Actual summary generation based on the meta-prompt is performed using two instruction-tuned Mixtral models (8x7B and 8x22B) in an Ollama deployment. These two models were selected because they support multilingual generation (including German), can be self-hosted for data-protection and non-disclosure reasons, and have been used in prior work on legal and administrative text processing [23, 24].

Batch Workflow for Similarity Benchmarking The experimental prototype as depicted in Fig. 1 supports automation of benchmark experiments based on the RAG pipeline (see Study 1 in Section 4.1). The data collection, including all JDs and each description's four components, can be programmatically retrieved from the database; different text variants can be generated (based on a comparison scenario); and a prompt template can be completed based on the preprocessed input text. The prompts were

expert-defined and followed a meta-prompting approach to ensure comparability between different models. The resulting texts and the selected reference texts can then be automatically compared using a battery of text-similarity computations (i.e., embedding-based, lexical, and n-gram similarity measures).

Interactive Human-in-the-Loop Workflow The pipeline is implemented in a Jupyter-based environment with a `voilà` Web user interface (see Fig. 1). Domain experts (appraisal officers) interact with the system through selection menus connected to the PostgreSQL database (Ⓐ in Fig. 1). This way, the compile the text corpus submitted to the generator. For each content type of job descriptions (tasks, goals, activities, requirements), in turn, the system retrieves relevant text corpora and generates a candidate summary (Ⓑ in Fig. 1). Appraisal officers then review the candidate summary for this content type and provide structured and unstructured feedback, Structured feedback is collected via a six-items questionnaire embedded in the interface (Ⓒ in Fig. 1). The questionnaire captures perceived quality in terms of language used, naturalness, clarity, factuality, non-discriminatibility, overall usefulness. In addition, domain experts can author an improved summary as unstructured feedback (using the generated text as prototype). This workflow is repeated (per case) four times, once for each content type. In addition, experts may re-iterate starting from loading different selectors. This forms the basis for Study 2 (see Section 4.2).

4. Multi-Study Design

The experiment prototype was employed as experimental material in two studies to answer the research questions: (1) in a benchmark study, selected text-similarity constructs were automatically contrasted using several predefined criteria and comparison scenarios; (2) in a field study, the perceived quality of generated texts by involving domain experts was systematically investigated. The results of the former study served as the evidence for preparing and designing the second, human-participant study.

Table 1
Similarity constructs used for summarization evaluation

Type	Constructs
Embedding-based (cosine)	`mixedbread-ai/deepset-mxbai-embed-de-large-v1`[1]; `bert-base-german-cased`[2]; `EuroBERT-2.1B`[3]; `EuroBERT-610M`[4]; `JobBERT`[5]; `Sentence-BERT (SBERT)`[21][6]
Lexical / statistical	TF–IDF [20] (grouped by section, department, and unit); BM25[25]
N-gram-based	Symmetric sentence-level BLEU [26]

4.1. Study 1: Comparing Similarity Constructs

The first study investigated the comparative performance of nine text-similarity constructs for capturing the semantic match between generated and reference texts (RQ_{1a}). Generated inputs were job-description summaries provided by our pipeline; once using a small (8x7B), once a large model (8X22B; see RQ_{1b}). As reference texts, derivative and predefined texts were used to implement different matching scenarios. A senior domain expert selected a subset of **53** representative JDs from the test dataset according to salary group, function, and harmonization goals. Similarity computation was performed based on three

[1] https://huggingface.co/mixedbread-ai/deepset-mxbai-embed-de-large-v1 (last accessed: March 17, 2026)
[2] https://huggingface.co/bert-base-german-cased (last accessed: March 17, 2026)
[3] https://huggingface.co/EuroBERT/EuroBERT-2.1B (last accessed: March 17, 2026)
[4] https://huggingface.co/EuroBERT/EuroBERT-610m (last accessed: March 17, 2026)
[5] https://huggingface.co/jjzha/jobbert-base-cased (last accessed: March 17, 2026)
[6] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2 (last accessed: April 8, 2026)

selected JD elements: **goals**, **tasks**, and **activities**. This is because these three elements are particularly relevant for the appraisal process. For each JD and each of the three components (goals, tasks, activities), we constructed five different textual variants:

text1 The first item (one goal, one task, one activity) from the original list of goals, tasks, or activities.
text2 A summary generated by the LLM using the complete list of items as input.
text3 A summary generated from the complete list of items *excluding* **text1**.
text4 A list of goals, tasks, or activities from a completely unrelated job domain (e.g., a lifeguard job description).
text5 A text completely unrelated to job descriptions (e.g., a children's bedtime story).

To compute measures of semantic matching between the different text variants, we computed pairwise similarity scores using embedding-based, lexical/statistical, and n-gram-based similarity constructs (Table 1). To assess the discriminative capabilities of the similarity measures more rigorously, we defined a three-class evaluation task based on characteristic comparison pairings of the five texts.

- **High-similarity class:** (text1, text2), (text2, text3), (text1, text3)
- **Medium-similarity class:** (text[1–3], text4)
- **Low-similarity class:** (text[1–3], text5)

This setup allows us to systematically investigate which similarity constructs measure (i) whether summaries preserve central semantic content from the source list, (ii) whether individual content items (tasks, activities, goals) are reflected in the generated summaries, and (iii) whether the content is domain-specific.

Across the constructed pairs, this resulted in 16 representative instances per JD: six high-, five medium-, and five low-similarity pairs. For each similarity construct, cosine similarity scores are computed and ranked. Instead of using absolute thresholds, we derive class predictions from rank positions to ensure comparability across models of different scales. Pairs ranked ≤ 6 are classified as high similarity, ranks 7–11 as medium similarity, and > 11 as low similarity. In addition, we trained a decision-tree classifier on the rank features to analyse confusion patterns across the three classes

Table 2 reports the classification accuracy for all similarity measures across tasks, goals, and activities. SBERT achieves the highest accuracy across all evaluated components, while lexical measures such as TF–IDF and BM25 obtain substantially lower scores. Table 3 presents confusion matrices from the rank-based classification analysis across all three JD elements. In all cases, low-similarity pairs are robustly separated, while most misclassifications occur between high- and medium-similarity classes.

Table 2
Three-class classification accuracy (%) across components.

Metric	Tasks	Goals	Activities
SBERT	96.4	86.9	77.4
mix_bread	81.5	82.4	67.8
bert_base	71.6	70.3	54.5
eurobert610M	68.9	68.8	66.3
jobBert	71.1	68.5	50.5
eurobert2.1B	50.3	61.8	57.3
bleu_symmetric	61.6	59.4	52.7
tfidf	53.9	54.4	49.2
bm25	26.5	27.3	28.2

To further investigate how similarity measures behave across content types, we inspected the univariate distribution of similarity scores for all pair combinations (see Figure 2). Embedding-based approaches generally show a clearer separation between in-domain and out-of-domain pairs than lexical baselines such as TF–IDF and BM25. In particular, low-similarity pairs involving unrelated texts (text5) are consistently assigned lower similarity scores by dense embedding models. However, the overlap between high- and medium-similarity pairs increases for more abstract components such as goals, indicating that the abstraction level influences semantic discriminability.

Table 3
Confusion matrices for the three-class classification analysis (rows: actual class, columns: predicted class).

	Tasks			Goals			Activities		
Actual	Low	Med	High	Low	Med	High	Low	Med	High
High	0	16	302	5	43	270	3	22	293
Medium	0	249	16	0	234	31	13	197	55
Low	318	0	0	318	0	0	314	2	2

Regarding the model size, both Mixtral models (`mixtral:8x22b`, `mixtral:8x7b`) produced semantically coherent summaries; there was no stable difference between the larger and the smaller model across JD elements.

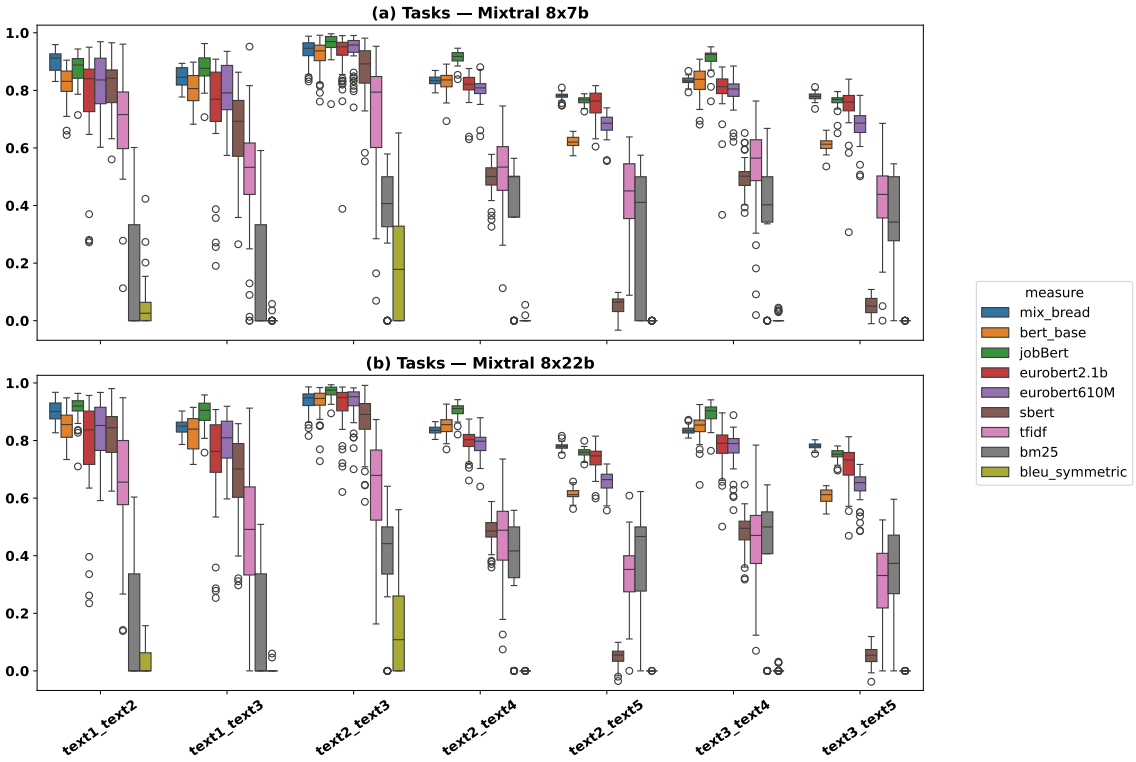

Figure 2: Univariate distributions of nine different similarity scores on task descriptions (*Aufgaben*) across seven comparison scenarios; summaries (text2, text3) generated using Mixtral 8x7B (top) and Mixtral 8x22B (bottom)

4.2. Study 2: Expert-Based Evaluation

The second study addressed how domain experts (appraisal officers) judge the quality characteristics of the generated JDs (RQ_2). The study was designed as a small-N randomized field experiment. Participants were invited to work on evaluation tasks on generated JDs in a natural, pre-existing work setting.

Planning & Design Participants were volunteers from a pool of 16 appraisal officers and therefore constituted a self-selection sample. Thirteen experts volunteered to participate in the study, of whom nine completed all three profile tasks. Participants were randomly assigned one of three predefined sequences of work tasks. The assignment was only known to the participants, and participants were asked not to reveal their assignment to researchers or fellow participants. The experimental material comprised three job profiles (A, B, C) predefined by a senior domain expert based on three characteristics that represent distinct administrative contexts and levels of responsibility: occupation group (*Verwendungsgruppe*), function group (*Funktionsgruppe*), and specific function (*Funktion*). Each profile

Table 4
Expert evaluation of generated job description components (median and interquartile range). Numbers next to profiles represent the number of responses. One participant did not complete Profile A.

Profile	Goals		Tasks		Activities		Requirements	
	MD	IQR	MD	IQR	MD	IQR	MD	IQR
A (8)	2.0	1.0	2.0	1.0	2.0	2.0	2.0	1.0
B (9)	2.0	1.0	2.0	1.0	2.0	1.0	3.0	1.0
C (9)	3.0	1.5	2.5	1.0	2.0	2.0	2.0	2.0

represented a group of approximately seven job descriptions in the test dataset. The work tasks of the participants were to operate the experimental prototype interactively to apply the selectors for each profile, trigger the generation of four text segments (tasks, goals, activities, requirements) of a new job description using the Mixtral 8x22B pipeline variant, and provide their evaluation of the generated texts. The experiment adopted a 3×3 Latin Square design based on the three job profiles A, B, and C to mitigate ordering (carry-over) effects. The three resulting profile sequences (ABC, BCA, and CBA) were assigned to the volunteer participants randomly and blinded. Quantitative and qualitative expert data were collected using a pre-questionnaire on individual AI literacy [27], a six-item questionnaire on the perceived quality of the generated job description (Table 5), and a free-text form to collect participants' text revisions (optional). The main feedback questionnaire consisted of six Likert-scale items with response options ranging from 1 ("strongly agree") to 5 ("strongly disagree"). The questions covered multiple qualitative aspects, including linguistic correctness, naturalness, comprehensibility, factual correctness, non-discrimination, and perceived usefulness (see Appendix A).

Execution Domain experts and project owners were invited to a briefing workshop in October 2025. The workshop included a demonstration of the experimental prototype, the study procedure, and materials (task description). The participants drew their profile sequences and were instructed on how to seek help and assistance. The pre-questionnaire on AI literacy was completed by all participants at the end of the workshop. Consent forms were distributed and collected. During the predefined two-week working period, 9 participants submitted their feedback containing the 12-item questionnaire and the free-text comment for each of the four components (goals, tasks, activities, and requirements) of a given profile. All three job profiles were assessed between 8 and 9 times. In a debriefing workshop, the results were presented to the participants and project owners in December 2025. In addition, group feedback on the experimental procedure and the prototype was gathered.

Analysis Results are reported using the median (MD) and interquartile range (IQR). Table 4 shows that tasks and activities are consistently rated positively across all profiles (MD = 2), indicating that operational content can be reliably generated. Goals and requirements exhibit greater response variability, with neutral ratings for goals in profile C and requirements in profile B, suggesting that more abstract or context-dependent components, such as goals and requirements, have a lower perceived quality. Table 5 reports linguistic and qualitative assessments. Language correctness and naturalness receive stable positive ratings across all profiles (MD = 2). In contrast, clarity, factual correctness, and usefulness show neutral medians for profiles B and C, reflecting increased sensitivity to domain-specific context. Non-discrimination is rated very positively across all profiles (MD = 1).

In summary, the expert evaluation indicates that the proposed RAG-based approach produces linguistically correct and largely acceptable job description components, particularly for tasks and activities.

5. Discussion

This section interprets the study findings in relation to the research questions, focusing on the reliability of similarity measures for evaluating LLM-generated job descriptions (RQ1) and on expert assessments of the generated content (RQ2).

With respect to **RQ1**, the benchmark study shows that embedding-based approaches provide a more

Table 5
Expert evaluation of linguistic and qualitative aspects of generated job descriptions (median and interquartile range). An asterisk (*) indicates that at least one expert assigned a disagreement rating (code ≥ 4), while a double asterisk (**) indicates that at least one expert assigned a strong disagreement rating (code ≥ 5).

Profile	Language		Naturalness		Clarity		Factual		Non-discr.		Usefulness	
	MD	IQR	MD	IQR	MD	IQR	MD	IQR	MD	IQR	MD	IQR
A (8)	2.0*	1.0	2.0*	1.5	2.0*	0.0	2.0*	1.0	1.0*	1.0	2.0*	2.0
B (9)	2.0*	1.0	2.0**	1.0	3.0*	2.0	3.0**	1.0	1.0*	1.0	3.0**	2.0
C (9)	2.0*	2.0	2.0*	2.0	3.0**	1.0	3.0*	1.0	1.0*	1.5	3.0*	2.0

reliable proxy for semantic alignment between generated and reference job description components than lexical baselines. Among the evaluated methods, Sentence-BERT (SBERT) achieves the highest rank-based classification accuracy across tasks, goals, and activities, consistently separating semantically unrelated content from related descriptions. In contrast, lexical similarity measures such as TF–IDF and BM25 perform substantially worse, indicating that surface-level term overlap is insufficient for capturing semantic equivalence in this domain.

The studies further reveal that semantic separability depends on the abstraction level of the JD component. Operational sections, such as tasks, achieve the highest classification accuracy, while goals and activities show greater overlap between similarity classes. The confusion matrices indicate that low-similarity pairs are reliably identified, whereas most misclassifications occur between high- and medium-similarity cases. This suggests that more abstract job description components allow for greater paraphrastic variation and conceptual interpretation, making precise semantic comparison more challenging.

Regarding the influence of model size, the automatic evaluation does not reveal a consistent advantage of the larger Mixtral 8x22B model over the 8x7B variant. However, the study was not designed as a comprehensive benchmark of LLM architectures. Instead, both models were used as representative locally deployable generators within the RAG workflow, allowing us to isolate the impact of similarity measures and evaluation strategies. Differences between similarity measures are substantially larger than differences between the two generation models. Although qualitative inspection indicates that the larger model occasionally produces more refined formulations, these improvements are not consistently reflected in rank-based evaluation results.

With respect to **RQ2**, the expert-based evaluation indicates that the generated job description components are generally perceived as linguistically correct and useful. Experts consistently rate operational sections such as tasks and activities positively, suggesting that these components can be summarized reliably using the proposed RAG-based workflow. In contrast, more abstract components such as goals and requirements show greater variability in expert assessments, reflecting their stronger dependence on contextual and organizational interpretation.

In summary, results indicate that embedding-based similarity measures provide an effective mechanism for evaluating semantic alignment in structured administrative texts, expert validation remains essential for assessing contextual adequacy and practical usefulness in more abstract JD components.

Limitations The dataset available (for both studies) is based on a small selection of documents compared to the entire existing corpus. This was because a critical share of JDs are not yet digitized, and others are stored in a decentralized manner. The dataset available is, hence, a convenience sample provided by our project partners to validate the usefulness of the experimental prototype. The selection of two LLMs (Mixtral) was justified at the time of study planning (Q4 2024) and aligns with the driving RQ of Study 1 (see Section 3). Still, our findings might not generalise to more recent LLMs reportedly used in HR domains (e.g., Gemma [24]). Study 2's limitations arise from the uncontrolled field-study setting and a comparatively small, self-selected sample of nine domain experts from a pool of 16. The mean AI literacy score [27] from 12 items across all four literacy dimensions (awareness, evaluation, usage, ethics) was comparatively high (mean±SD): 5.1±.77 (7-point Likert). In the absence of comparable evidence in this domain, and beyond, the relative proficiency of our small pool cannot be concluded.

Interestingly, within the pool, we found that the more proficient a participant was, the fewer of the three tasks (profiles), if any, were actually tackled and submitted. All else equal, every additional point of literacy decreased the odds of task fulfilment by 46.2% to 99.8% (95%CI).

6. Related Work

RAG has become a prominent strategy for improving factual consistency and traceability in LLM outputs. Lewis et. al. [13] introduced RAG as a framework combining retrieval with generative models for knowledge-intensive NLP tasks, demonstrating improved performance. Subsequent work has explored retrieval-enhanced generation and grounding mechanisms [28], while recent surveys highlight the role of retrieval in mitigating hallucinations and improving traceability [17, 29]. In regulated administrative contexts, where accountability and compliance are critical, purely generative approaches remain insufficient. Human-centered AI research therefore stresses the importance of human-in-the-loop validation in high-stakes settings [30]. Our evaluation framework therefore combines automatic similarity analysis with structured expert assessment for administrative documents.

Evaluating LLMs in knowledge-intensive settings remains challenging. Foundation model studies caution against relying solely on surface-level metrics [31], and benchmarks such as KILT [15] focus mainly on question answering rather than structured document generation. In contrast, we assess generated job description components using both semantic similarity measures and expert evaluation, addressing the evaluation needs of regulated domains.

Semantic similarity modeling plays a central role in assessing generated content. Contextual embeddings such as BERT [16] and Sentence-BERT [21] provide dense semantic representations for sentence-level comparison, while benchmarking studies such as BEIR [32] indicate the domain sensitivity of embedding models. Traditional lexical and statistical measures, including TF–IDF [20], BM25 [25], and BLEU [33] emphasize surface-level term overlap. Prior work in summarization evaluation has shown that overlap-based metrics can fail to capture semantic abstraction and factual consistency [34], motivating the use of embedding-based approaches for deeper semantic comparison.

In the public administration and labor market domain, prior work has highlighted the need for standardized occupational taxonomies and semantic interoperability, including ESCO [8] and ontology-based skills matching approaches [7, 35]. However, empirical studies evaluating LLM-assisted generation of structured job description components remain scarce, particularly in regulated administrative contexts.

7. Conclusions and Future Work

This paper investigated the use of RAG to support the semi-automatic generation and harmonization of structured job description components in public administration. By combining embedding-based retrieval, controlled LLM prompting, and expert validation, the proposed pipeline enables the generation of standardized summaries grounded in existing administrative documents.

The evaluation combined automatic similarity analysis with structured expert assessment to examine both semantic alignment and perceived output quality. The results show that embedding-based similarity measures, particularly Sentence-BERT, provide a reliable mechanism for evaluating semantic correspondence between generated and reference job description components. At the same time, expert assessments confirm that the generated outputs are linguistically sound and largely acceptable for practical use, especially for operational components such as tasks and activities.

The findings also highlight that the abstraction level plays an important role in evaluating generated content. While operational job description elements can be summarized reliably, more abstract components, such as goals and requirements, remain sensitive to contextual interpretation and therefore benefit from expert oversight.

Overall, the study demonstrates that RAG-based workflows can support knowledge-intensive administrative tasks by assisting experts in drafting and harmonizing structured job descriptions while maintaining human control over context-sensitive decisions.

Future work involves extending the dataset to include more heterogeneous JDs. This, in turn, will require exploring a refined retrieval method (e.g., by preprocessing JDs elements using hierarchical text classification) to generate representative, yet size-limited text corpora to avoid context overflows during generation. In addition, a second test dataset will be mined based on a comparable, publicly available data source: public-sector job announcements from the Austrian Federal Job Portal.[2] Job announcements and job descriptions share common structural elements (e.g., activities and tasks) and a similar legal, as well as domain (HR) expression.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT 5.2 in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

Acknowledgments

This research project was partly funded by the Public Service and Administrative Innovation Section of the Austrian Federal Chancellery. We thank Andrea Graf-Langheinz for her project contributions and the participants of the field study for their time and constructive feedback.

References

[1] N. Edelmann, I. Mergel, T. Lampoltshammer, Competences that foster digital transformation of public administrations: an Austrian case study, Administrative Sciences 13 (2023) 44. doi:https://doi.org/10.3390/admsci13020044.

[2] A. V. Switasarra, R. D. Astanti, Literature review of job description: Meta-analysis, International Journal of Industrial Engineering and Engineering Management 3 (2021) 33–41. doi:10.24002/ijieem.v3i1.4923.

[3] R. Stock-Homburg, M. Groß, Personalmanagement: Theorien–Konzepte–Instrumente, Springer-Verlag, 2019.

[4] Austrian Court of Audit, Personalbewirtschaftung des bundes mit dem schwerpunkt arbeitsplatzbewertungen, 2022. URL: https://www.rechnungshof.gv.at/rh/home/home/Bund_Personalbewirtschaftung_des_Bundes_mit_dem_Schwerpunkt_.pdf.

[5] B. E. Becker, M. A. Huselid, Strategic human resources management: where do we go from here?, Journal of management 32 (2006) 898–925.

[6] Federal Chancellery Austria, Arbeitsunterlage zum Erstellen von Arbeitsplatzbeschreibungen, 2023. URL: https://oeffentlicherdienst.gv.at/wp-content/uploads/2023/01/Arbeitsunterlage-zum-Erstellen-von-Arbeitsplatzbeschreibungen.pdf.

[7] S. Sreerambatla, Synthesizing Job Market Data: Building a Unified Repository Using ONET and ESCO, International Journal of Science and Research 9 (2020). doi:10.21275/SR24628183641.

[8] J. D. Smedt, M. le Vrang, A. Papantoniou, ESCO: Towards a Semantic Web for the European Labor Market, in: Workshop Proc.Linked Data on the Web (LDOW 2015), volume 1409 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2015. URL: https://ceur-ws.org/Vol-1409/paper-10.pdf.

[9] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, The Journal of Machine Learning Research 21 (2020) 5485–5551. doi:10.5555/3455716.3455856.

[10] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language Models are Few-Shot Learners, Advances in Neural Information

[2] https://bund.jobboerse.gv.at/sap/bc/jobs/, last accessed: March 16th, 2026

Processing Systems 33 (2020) 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[11] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[12] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Computing Surveys 55 (2023) 1–38. doi:10.1145/3571730.

[13] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, Advances in Neural Information Processing Systems 33 (2020) 9459–9474. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

[14] Federal Chancellery Austria, Grundlagen für die Arbeitsplatzbewertung, 2023. URL: https://oeffentlicherdienst.gv.at/wp-content/uploads/2023/01/Grundlagen-fuer-die-Arbeitsplatzbewertung.pdf.

[15] F. Petroni, A. Piktus, A. Fan, P. Lewis, M. Yazdani, N. De Cao, J. Thorne, Y. Jernite, V. Karpukhin, J. Maillard, et al., Kilt: a benchmark for knowledge intensive language tasks, in: Proc. 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, ACL, 2021, pp. 2523–2544. doi:10.18653/v1/2021.naacl-main.200.

[16] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1, ACL, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[17] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, et al., Retrieval-augmented generation for large language models: A survey, arXiv preprint arXiv:2312.10997 2 (2023) 32.

[18] I. Mani, Automatic summarization, John Benjamins Publishing Company, 2001.

[19] A. See, P. J. Liu, C. D. Manning, Get to the point: Summarization with pointer-generator networks, in: Proc. 55th Annual Meeting of the Association for Computational Linguistics, volume 1, ACL, 2017, pp. 1073–1083. doi:10.18653/v1/P17-1099.

[20] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Information processing & management 24 (1988) 513–523. doi:10.1016/0306-4573(88)90021-0.

[21] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in: Proc. 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, ACL, 2019, pp. 3982–3992. doi:10.18653/v1/D19-1410.

[22] Y. Zhang, Y. Yuan, A. C.-C. Yao, Meta Prompting for AI Systems, 2025. URL: https://arxiv.org/abs/2311.11482. arXiv:2311.11482.

[23] J. Breton, M. M. Billami, M. Chevalier, H. T. Nguyen, K. Satoh, C. Trojahn, M. M. Zin, Leveraging LLMs for legal terms extraction with limited annotated data, Artificial Intelligence and Law (2025) 1–27. doi:10.1007/s10506-025-09448-8.

[24] T. K. Dasaklis, P. G. Giannopoulos, D. Koutras, V. Malamas, P. Chountalas, Large Language Models in Human Resource Management: a systematic literature review of applications, open issues and future research directions, Available at SSRN 5314976 (2025). doi:10.2139/ssrn.5314976.

[25] S. E. Robertson, K. S. Jones, Relevance weighting of search terms, Journal of the American Society for Information science 27 (1976) 129–146. doi:10.1002/asi.4630270302.

[26] B. Chen, C. Cherry, A Systematic Comparison of Smoothing Techniques for Sentence-Level BLEU, in: Proc. 9th Workshop on Statistical Machine Translation, ACL, 2014, pp. 362–367. doi:10.3115/v1/W14-3346.

[27] B. Wang, P.-L. P. Rau, T. Yuan, Measuring user competence in using artificial intelligence: validity and reliability of artificial intelligence literacy scale, Behaviour & Information Technology 42 (2023) 1324–1337. doi:10.1080/0144929X.2022.2072768.

[28] G. Izacard, E. Grave, Leveraging passage retrieval with generative models for open domain

question answering, in: Proc. 16th Conference of the European Chapter of the Association for Computational Linguistics, ACL, 2021, pp. 874–880. doi:10.18653/v1/2021.eacl-main.74.
[29] G. Mialon, R. Dessì, M. Lomeli, C. Nalmpantis, R. Pasunuru, R. Raileanu, B. Rozière, T. Schick, J. Dwivedi-Yu, A. Celikyilmaz, E. Grave, Y. LeCun, T. Scialom, Augmented Language Models: a Survey, Transactions on Machine Learning Research (2023). doi:10.48550/arXiv.2302.07842.
[30] B. Shneiderman, Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy, International Journal of Human–Computer Interaction 36 (2020) 495–504. doi:10.1080/10447318.2020.1741118.
[31] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al., On the opportunities and risks of foundation models, arXiv preprint arXiv:2108.07258 (2021).
[32] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych, BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models, in: Proc. Neural Information Processing Systems Track on Datasets and Benchmarks, 2021. doi:10.48550/arXiv.2104.08663.
[33] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a Method for Automatic Evaluation of Machine Translation, in: Proc. 40th Annual Meeting of the Association for Computational Linguistics, ACL, 2002, pp. 311–318. doi:10.3115/1073083.1073135.
[34] J. Maynez, S. Narayan, B. Bohnet, R. McDonald, On Faithfulness and Factuality in Abstractive Summarization, in: Proc. 58th Annual Meeting of the Association for Computational Linguistics, ACL, 2020, pp. 1906–1919. doi:10.18653/v1/2020.acl-main.173.
[35] A. C. Tuan, M. T. Dang, H. N. Do, V. K. Solanki, J. Torres, R. Gonzalez Crespo, T. N. A. Nguyen, Ontology and its applications in skills matching in job recruitment, Applied Ontology 19 (2024) 287–306. doi:10.3233/AO-240019.

Appendix

A. Expert Evaluation Questionnaire

This appendix lists the questionnaire used in the expert-based evaluation described in Section 4.2. To improve transparency and reproducibility, we provide both the original German wording and an English translation of each item (Table 6).

Table 6
Expert evaluation questionnaire items

German (original)	English translation
Der Text ist sprachlich korrekt (Rechtschreibung, Grammatik, idiomatische Ausdrucksweise).	The text is linguistically correct (spelling, grammar, and idiomatic expression).
Der Text klingt so, als wäre er von einer deutschen Muttersprachler:in verfasst.	The text reads as if it were written by a native German speaker.
Die Aussagen sind eindeutig und für den vorgesehenen Personenkreis verständlich.	The statements are clear and understandable for the intended audience.
Die Inhalte sind fachlich richtig.	The content is factually correct.
Der Text ist frei von diskriminierenden Formulierungen.	The text does not contain discriminatory language.
Der Text ist für die Erstellung bzw. Überarbeitung von Arbeitsplatzbeschreibungen nützlich.	The text is useful for creating or revising job descriptions.

B. Additional Similarity Distributions

This appendix provides additional similarity distribution plots for the goal and activity components used in Study 1. While the main paper presents the task component as a representative example, the plots on Figure 3 illustrate how similarity measures behave for goals and activities.

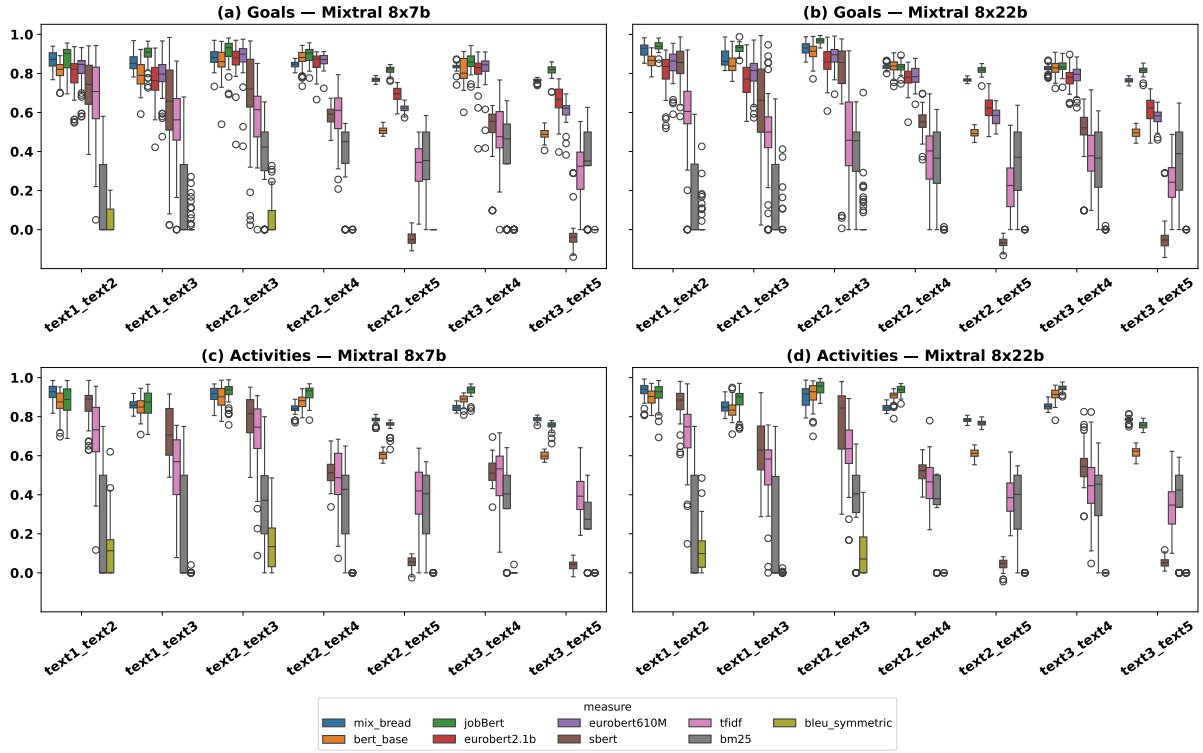

Figure 3: Similarity distributions across measures for job description components. Top row: goals (*Ziele*); bottom row: activities (*Taetigkeiten*). Left: Mixtral 8×7B; right: Mixtral 8x22B.

C. Prompt Template Used for JD Component Generation

The prompt template used in the RAG pipeline (Section 3) to generate job description components was developed with domain experts to ensure consistent structure, style, and content. For reproducibility, it is presented in a generalized form with placeholders: CONTENT_TYPE (e.g., *Aufgaben*, *Ziele*, *Tätigkeiten*, *Anforderungen*) and INPUT_TEXT (the retrieved JD text). It also enforces constraints such as language consistency, length limits, removal of redundancy, and avoidance of filler phrases.

```
Antworte ausschließlich auf Deutsch.
Erstellen Sie eine Liste von {CONTENT_TYPE} für eine offene Stelle
auf der Grundlage der nachstehenden Beschreibungen und Beispiele:

Listenformat: Die zusammengefassten {CONTENT_TYPE} müssen durch „;” getrennt sein.
Maximale Länge: Die Antwort darf höchstens 200 Wörter enthalten.
Nur Hauptinhalte: Entfernen Sie irrelevante Details, Zusatzinformationen
und Kontextbeschreibungen.
Vermeidung von Füllwörtern: Verwenden Sie keine Konnektoren oder
verbindende Phrasen wie „Des Weiteren", „Außerdem" oder „Schließlich".
Keine Wiederholungen: Doppelte oder redundant ähnliche Inhalte
sind zu vermeiden.
Nur Fachsprache: Verwenden Sie klare, sachliche Begriffe, ohne
Ausschmückungen oder allgemeine Formulierungen.
Sprache: Die Antwort muss in DEUTSCHER Sprache verfasst sein.
Leiten Sie die Zusammenfassung nicht ein mit z.B. „Zusammenfassung:",
„Die {CONTENT_TYPE} umfassen:".
Beginnen Sie die Aufzählung nicht mit einer Einleitung wie z.B.:
„Zusammenfassung:", „Die {CONTENT_TYPE} umfassen:" oder „{CONTENT_TYPE}".

{CONTENT_TYPE}: {INPUT_TEXT}
```