Improving Online Job Advertisement Analysis via Compositional Entity Extraction

Anonymous ACL submission

Abstract

We propose a compositional entity modeling framework for requirement extraction from online job advertisements (OJAs). To more accurately capture the structure of requirements in OJAs, we reframe the task from identifying single-span annotations to modeling complex, tree-like structures that connect atomic entity types via typed relationships. Based on this schema, we introduce GOJA, a high-quality dataset of 500 German job ads. GOJA captures the internal semantics of job requirements, including roles, tools, experience levels, attitudes, and their functional context.

We describe the annotation process, report strong inter-annotator agreement, and benchmark transformer models to demonstrate the feasibility of training on this structure. To illustrate the analytical potential of our approach, we present a focused case study on AI-related job requirements. We show how our proposed compositional representation enables new types of labor market analyses.

1 Introduction

Online Job Advertisements (OJAs) serve as a critical data source for understanding labor market dynamics across disciplines such as labor market research, education, and human resources (Khaouja et al., 2021). They offer detailed and up-to-date insights into in-demand skills, required qualifications, and evolving industry trends. By analyzing OJAs, researchers can identify skill gaps and inform educational planning (Lima et al., 2018; Giabelli et al., 2021; Buchmann et al., 2022; Atalay et al., 2020, 2023). Job Ads have also been used in recruiting research (Castilla and Rho, 2023; Kim and Angnakoon, 2016) and for developing job recommendation systems via cv matching (Ntioudis et al., 2022; Smith et al., 2021; Belloum et al., 2019).

Work on Information Extraction (IE) in OJAs has mostly focused on skills extraction (see survey by Senger et al., 2024). Work extracting other information includes job tasks (Atalay et al., 2018, 2020, 2023), job titles (Baskaran and Müller, 2023; Li et al.; Giabelli et al., 2021; Rahhal et al., 2023), work tools (Güntürk-Kuhl et al.) and formal qualifications (Brown and Souto-Otero, 2020; Müller; Schimke, 2023; Börner et al., 2018). Collectively, these entities can be summarized as *requirements*, reflecting aspects of the position sought that pertain to the candidate.

Limitations of single-span requirement modeling. Most existing approaches to requirement extraction in OJAs rely on flat, span-based annotation schemes that treat expressions such as "Python", "ML", or "Previous work experience" as standalone entities. However, such representations fail to capture internal structure and logical relations.

Figure 1 illustrates this using three example sentences from a job ad. Each sentence is annotated with span-based baselines (top) and our framework (bottom).

In the first sentence, single-span schemes tend to annotate almost the entire sentence as a single span, since they cannot represent semantic links—such as the relation between *apply* and *machine learning algorithms*. This leads to semantically overloaded spans, as the difference between applying and, for instance, developing or managing ML systems cannot be made explicit otherwise. Long spans also not only increase ambiguity and model error rates (Zhang et al., 2022b), but also struggle to represent embedded or conjoined elements (Nguyen et al., 2024).

In the second sentence, "Python or Java" explicitly states these two programming languages as alternative requirements. Current approaches, however, mark both terms as independent skills, thus losing the disjunctive meaning. In addition, the associated experience level ("familiarity") is not modeled as part of the skill expression. In the third sentence, on the ohter hand, Green et al. (2022)

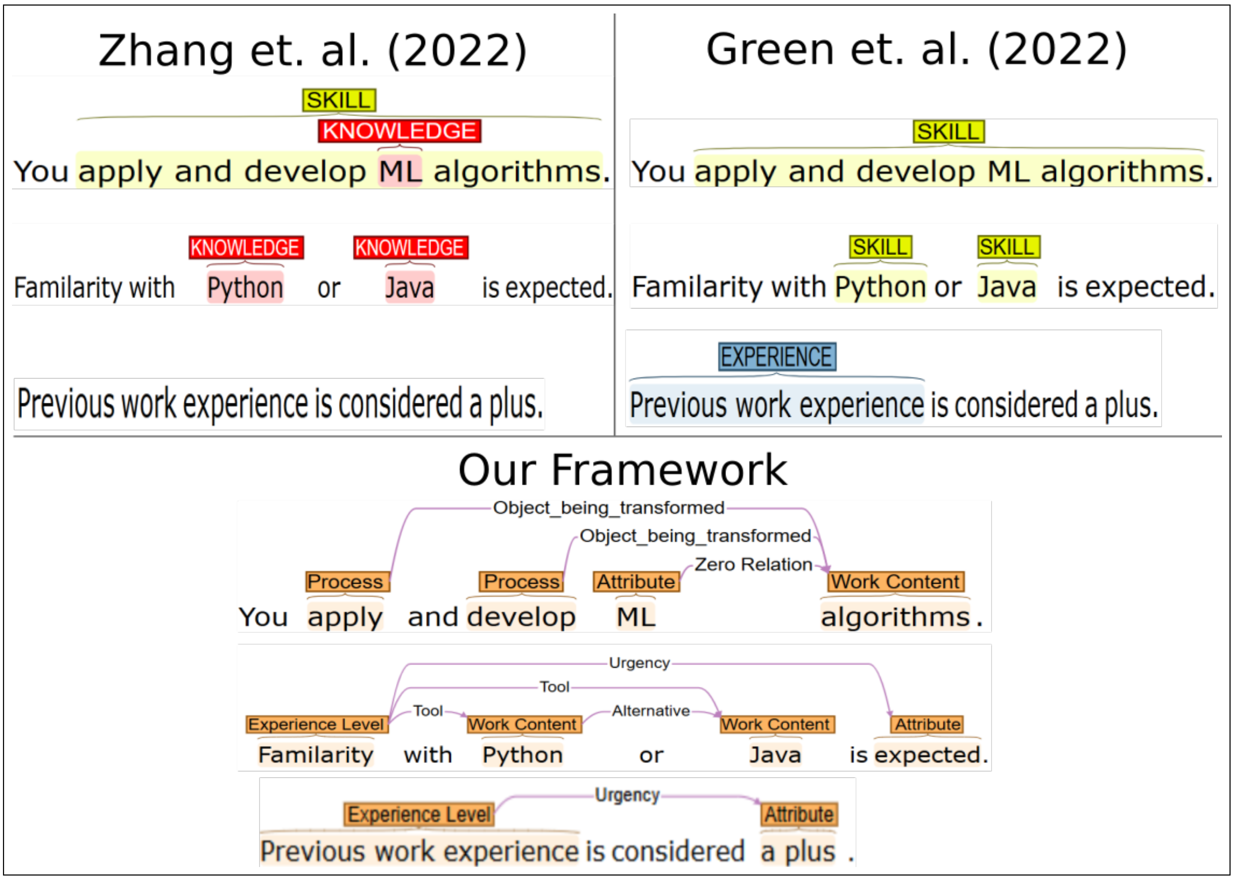

Figure 1: Side-by-side comparison of the same three sentences annotated via different requirement modeling approaches. For (Green et al., 2022) and (Zhang et al., 2022b), we annotated the sentences using their public annotation guidelines.

annotate "work experience" as a requirement only when it appears in isolation. When embedded in a more complex construction, such as being linked to a specific skill, it remains unannotated.

Finally, expressions that indicate the urgency or desirability of a requirement—such as "is expected" or "is a plus" in sentences 2 and 3—are, to our knowledge, not explicitly annotated in existing schemes. Yet such phrases carry critical semantic information.

Contributions. To address these challenges, we propose a compositional entity modeling framework that decomposes requirement descriptions into their constituent components and explicitly models their relationships. Consequently, we methodologically extend the entity extraction setup by additionally modeling typed relations between entities, enabling a structured representation of requirement expressions.

In more detail, our contributions are:

- We propose a compositional framework for modeling job requirements in OJAs, addressing limitations of single-span entity extraction by modeling entities and their relationships.
- We introduce GOJA[1], a manually annotated gold-standard dataset of 500 German job advertisements, containing over 22,000 entities and 13,000 typed relations.
- We demonstrate the feasibility and analytical value of our approach through (i) descriptive analyses of structural patterns in the data, (ii) benchmark experiments using transformer-based models for entity and relation extraction, and (iii) a focused case study on AI-related requirements.

2 Compositional Annotation of Job Advertisements: The GOJA Dataset

This section introduces GOJA. We first review related datasets in the area of requirement extraction

[1]We release the created GOJA to the research community upon acceptance of the paper.

Entity Type	Description	Example
Attitude	Indicates traits or dispositions desired in candidates.	You are adaptable
Attribute	Provides additional specifications about other entities.	You design logos for our customer
Experience Level	Indicates the level of knowledge or skills required.	Experience in Python
Formal Qualification	Identifies certifications or official qualifications required.	Bachelor's degree in Economics
Industry	Defines the industry or sector associated with the job.	You bring relevant experience in the automotive industry
Occupation	Specifies the role or position advertised.	We looking for a baker (m/f/d)
Process	Represents actions or sequences required to perform tasks.	You design Logos
Work Content	Describes the object or tool related to a task.	You design logos
Relation Type	**Description**	**Example**
Alternative	Denotes alternatives between entities.	Bachelor's degree or minimum of three years professional experience)
Coordination	Connects coordinated morphems within sentences.	You pre- and post- process texts.
Degree of Autonomy	Specifies the level of autonomy in task execution.	You help your supervisor prepare presentations
Detail	Illustrates subcategories or specifics of an entity.	You are experienced with at least one programming language like Python
Negation	Highlights excluded processes or tasks.	This role does not include care duties.
Object Being Transformed (OBT)	Links processes to the items or entities they affect.	You design new logos
Related Entity Parts (REP)	Links separated parts of an entity.	You set the annual budget up
Specialization	Adds specificity to qualifications or roles.	A Bachelor's degree in Economics
Tool	Connects processes to the tools or methods used.	You design logos using Illustrator
Urgency	Indicates the importance or necessity of an entity.	Experience in Python is mandatory
Zero Relation	Used where the relation is self-evident.	You bring experience in programming

Table 1: Overview of entity and relation types in our proposed annotation scheme. For relation types, the examples underline the subject and object entity of the respective relation.

from job advertisements, then describe our annotation schema, and finally detail the annotation process and resulting dataset statistics.

2.1 Related Datasets

We focus here on publicly available datasets for requirement extraction from online job advertisements. We restrict our scope to methodologically relevant datasets used for training or evaluating information extraction models — excluding purely analytical corpora (like in Atalay et al. (2020))

Despite the growing interest in this field, dataset availability remains limited. According to an overview provided by Zhang et al. (2022b), more than 80% of skill extraction studies do not release their datasets or annotation guidelines. To the best of our knowledge, no publicly available datasets exist for other requirement types such as job tasks, job titles, or formal qualifications.

A recent survey by Senger et al. (2024) summarizes the current landscape of skill-related datasets. Datasets released to the public are: **SAYFULLINA** (Sayfullina et al., 2018) presents an English dataset of soft skills, annotated via crowdsourcing using a predefined list and binary relevance labels. **GREEN** (Green et al., 2022) crowdsource both hard and soft skills in English ads, additionally labeling occupations, experience levels, and qualification indicators. **SKILLSPAN** (Zhang et al., 2022b) introduces expert-annotated spans for both skills and knowledge concepts. **KOMPETENCER** (Zhang et al., 2022a) provides Danish span-level annotations aligned with the ESCO taxonomy, covering both coarse and fine-grained skill labels. **DECORTE** (Decorte et al., 2022) offers Dutch skill annotations manually mapped to ESCO

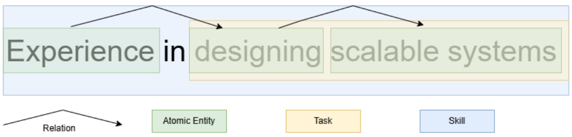

Figure 2: Example of analysis chains for skills and tasks.

concepts, serving as gold-standard data for evaluation. **GNEHM-ICT** (Gnehm et al., 2022a) focuses on Swiss German ICT job ads, annotating related entities. **BHOLA** (Bhola et al., 2020) approaches the task differently, using document-level multi-label classification of English job ads based on a predefined skill inventory. **FIJO** (Beauchemin et al., 2022) provides French span-level skill annotations using sequence labeling. Skills are categorized into four predefined types—“Thoughts”, “Results”, “Relational”, and “Personal”—derived from public and proprietary taxonomies.

2.2 Proposed Annotation Schema

The key observation underlying our approach is that fuzzy concepts such as skills and tasks are often not directly represented in text as discrete, self-contained entities. Instead, they emerge compositionally from smaller, interrelated components. Our framework formalizes this by analyzing skills and tasks as chains of atomic entities linked by relations.

Table 1 provides a full overview of all 8 entity and 11 relation types in our annotation framework.

Tasks. Tasks are demand-side job elements that transform inputs into outputs within an economic context (Autor and Handel, 2013; Rodrigues et al., 2021). They can be described at varying levels of granularity. In our schema, the PROCESS entity captures the action, and the WORK CONTENT entity specifies its target or context. These are linked via relations that express semantic dependencies. Depending on its role, WORK CONTENT may refer to an OBJECT BEING TRANSFORMED (OBT)—e.g., a thing, concept, person—or to a *work tool* used to carry out the process (Fana et al., 2023)

Skills. Skills are defined as the ability to perform a task effectively (Rodrigues et al., 2021), representing the supply side of labor. In our framework, skills are modeled as tasks augmented by EXPERIENCE LEVEL entities. Figure 2 shows how the task "designing scalable systems" plus the entity "Experience" form a skill. This skill-task distinction underscores the importance of compositional modeling in capturing not just the components of tasks and skills but also their contextual modifiers. In this conceptualization, tasks entail certain skills but not vice versa.

Attitudes. Traits often labeled as *soft skills* are represented as ATTITUDE entities in our schema. Attitudes are psychological, emotional, or behavioral predispositions—e.g., empathy, adaptability, or stress tolerance—that support effective task performance (Rodrigues et al., 2021). Unlike skills, which are tied to specific tasks, attitudes pertain to broader domains of competence.

Other entities and relations. The other entities and relations have been derived inductively during annotation guideline development (see Section 2.3) based on the goals of our framework (e.g., FORMAL QUALIFICATION was introduced because we were interested in degrees mentioned), their frequent occurrence in patterns (e.g. URGENCY) or the need to correctly represent the meaning of the text (e.g. syntactically motivated relations like COORDINATION or REP). The most arbitrary categories are ATTRIBUTES and ZERO RELATION. Attributes provide additional context that may or may not be relevant for the analysis. They cannot stand alone, but specify details about primary entities. While Attributes may span longer phrases, all other entity types are defined as concisely as possible to balance annotation consistency and model performance. TThis design reduces complexity for key entities while capturing optional nuances through attributes as a flexible catch-all for contextual details. The ZERO RELATION applies to entities whose connection is self-evident and needs no further specification.

2.3 Dataset Annotation

To prepare a suitable dataset for annotation, we sampled 500 German job ads from Textkernel’s Jobfeed corpus, restricting to regular employment (excluding apprenticeships). A multivariate sampling approach balanced multiple factors (year of publishing, website source, WZ08 activity, ISCO08 occupation, contract type, and text length), aiming to minimize selection bias.

We conducted the annotation in two phases: (1) iterative guideline development and (2) final annotation of 500 OJAs:

Phase 1 Following Reiter et al. (2019), four original annotators (A) refined the guidelines over six rounds on small samples, comparing annotations and adjusting rules to ensure consistency and construct validity.

Phase 2 In the final phase (2), 15 researchers (A plus newly trained annotators B) participated. Group B received tutorials and performed test annotations; only those surpassing Krippendorff's $\alpha \geq 0.7$ proceeded. Each OJA was then double-annotated and curated by a third annotator (A). This yielded Krippendorff's $\alpha = 0.88$ for entities and $\alpha = 0.80$ for relations — values considered reliable by Krippendorff (2018).

Comparing our metrics to other work in the field, Green et al. (2022) report Cohen's $\kappa = 0.49$ and Krippendorff's $\alpha = 0.55$, while Zhang et al. (2022b) report Fleiss' κ between 0.70 and 0.75. Although the scores are not directly comparable due to differences in annotation schemes and task definitions, our results indicate a relatively high inter-annotator reliability.

2.4 Describing GOJA

Following the annotation process, we compiled the resulting data we refer to as GOJA ("German Online Job Advertisements"). GOJA yields 22,506 entities and 13,324 relations across 500 German-language OJAs. In this section, we provide an overview of key dataset properties and highlight compositional patterns that reflect the complexity of requirement expressions in real-world OJAs. Given our multivariate sampling approach, this distribution should approximate their occurrence in larger datasets. Figure 3 illustrates the distribution of key analytical units—tasks, skills, and attitudes—per document, as derived from the chains described in Section 2.2.

Explicit distinction between tasks and skills. Notably, concepts that are extracted as skills in other studies tend to be formulated as tasks in our conceptualization. This observation reflects how most analyses with OJA data (implicitly) equate job tasks with skills, i.e. the proficiency in these tasks. However, as employer-provided training is almost ubiquitous in Germany, especially in entry-level jobs in Germany (Lukowski et al., 2021), candidates are not expected to master all tasks at the outset. Consequently, our findings indicate that research could benefit from investigating why certain tasks are explicitly associated with an experience level while others are not.

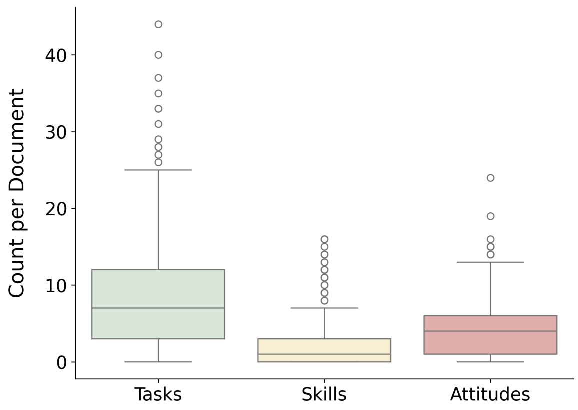

Figure 3: Boxplot showing the distributions of Skills, Tasks and Attitudes per document.

Comparing the frequency of skills and attitudes, it can be derived that in terms of typical OJA text zones (Gnehm, 2018; Gnehm and Clematide, 2020), our analysis reveals that skill segments in job advertisements predominantly consist of attitudes rather than hard skills.

High frequency of conjoined skills and tasks. Our analysis reveals that conjoined requirement structures are common in OJAs. A substantial share of both tasks (44%) and skills (30%) involve multiple linked components, such as one process affecting several work contents, or one experience level modifying several tasks or tools. These patterns occur more frequently than previously reported in comparable studies (Nguyen et al., 2024) and highlight the importance of explicitly modeling such structures. A more detailed breakdown of conjoined configurations is provided in Appendix A.

3 Applying GOJA

To demonstrate the practical utility of GOJA, we apply it in two ways. First, we train baseline extraction models to show that the compositional schema can be learned by transformer-based architectures. Second, we use these models to analyze AI-related requirements in a larger corpus of job ads, illustrating the analytical benefits of structured, relation-based modeling.

3.1 Baseline Models

To assess whether the GOJA annotation schema can be learned effectively, we train transformer-based models for both entity and relation extraction. These models form the basis for downstream applications and enable automated large-scale analysis.

3.1.1 Model Setup

We fine-tune four different pre-trained transformer models: German BERT (Devlin et al., 2019), German DistilBERT (Sanh, 2019), jobBERT-de (Gnehm et al., 2022b)—a variant of German BERT fine-tuned on German OJA data—and the multilingual XLM-RoBERTa (Conneau, 2019). For entity extraction, we use a token classification head on top of the pre-trained models.

For relation classification, we adopt a simple yet effective approach: Entities participating in a relation are marked with special tokens [E] and [/E] within their sentence, and the modified sequence is passed to a transformer-based sequence classification model. To handle candidate entity selection efficiently, we use a context window of four sentences, based on internal analyses, to determine potential entity pairs. Additionally, we introduce a NO RELATION class to distinguish entity pairs that do not share a relation. Since this results in a class imbalance, we randomly downsample the No Relation class to match the total number of instances in the other relation classes.

Prior to cross-validation, we determined suitable hyperparameters via grid search to optimize model performance. We report the F1-score averaged over five-fold cross-validation, ensuring robustness across different data splits. The dataset follows a 70-15-15 split into training, validation, and test sets, with all reported F1-scores computed exclusively on the unseen test set to provide a realistic assessment of generalization performance.

3.1.2 Performance Overview

Our experimental results are summarized in Table 2. We observe that XLM-RoBERTa clearly outperforms the other three models in both entity extraction and relation classification. Notably, jobBERT-de also achieves solid performance, improving over German BERT and German DistilBERT in both tasks. An interesting finding is that the performance gap among models is much larger in the entity subtask than in relation classification.

Model	Entity F1	Relation F1
German BERT	0.665 $\pm$ 0.025	0.836 $\pm$ 0.008
German DistilBERT	0.517 $\pm$ 0.024	0.788 $\pm$ 0.012
jobBERT-de	0.718 $\pm$ 0.013	0.874 $\pm$ 0.014
XLM-RoBERTa	**0.856** $\pm$ 0.012	**0.911** $\pm$ 0.007

Table 2: F1 scores and standard deviation for entity extraction and relation classification, averaged over five-fold cross-validation.

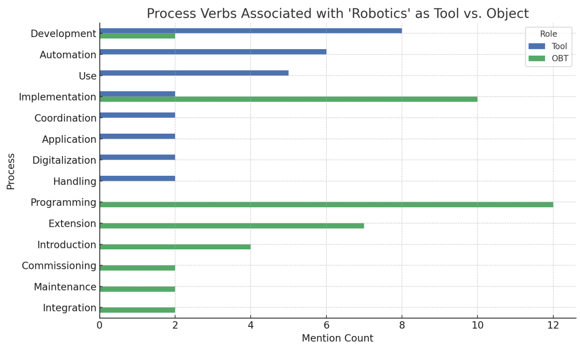

Figure 4: Process verbs associated with robotics as `Work Tool` vs. `Object Being Transformed` (OBT).

3.2 Case Study: Analyzing AI-related Requirements

To illustrate the analytical potential of our schema, we analyze OJAs mentioning Artificial Intelligence (AI). AI-related requirements are of growing interest in labor market research. From a corpus of 2.8 million ads, we selected approximately 19,000 matching a curated keyword list derived from a computer science ontology (Salatino et al., 2018) and a public repository (Peede and Stops, 2024). These ads were processed with our best-performing models (cf. Table 2), resulting in around 1.9 million entities and 1.9 million relations. In the following analysis, we examine AI-related entities and their relation chains to highlight structured patterns in job requirement descriptions.

Robotics as Tool and Object. The most central differentiation in job tasks in our framework lies in the relations `OBT` and `Work Tool` between `Work Content` and `Process` entities. Figure 4 shows the process verbs most frequently associated with keywords in *robotics* in each role, aggregated across verb variants. When labeled as a `Work Tool`, robotics appears in the context of operational actions such as *use*, *automation*, or *implementation*. In contrast, robotics as an `OBT` is associated with development-oriented verbs such as *programming*, *integration*, or *commissioning*. These findings highlight the advantage of contextualizing `Process` and `Work Content` relations to more accurately capture competence profiles. This distinguishes, for instance, between operational usage and developmental expertise.

Occupational Framing of Machine Learning. To further demonstrate the analytical value of our schema, we examine how the term *machine learn-*

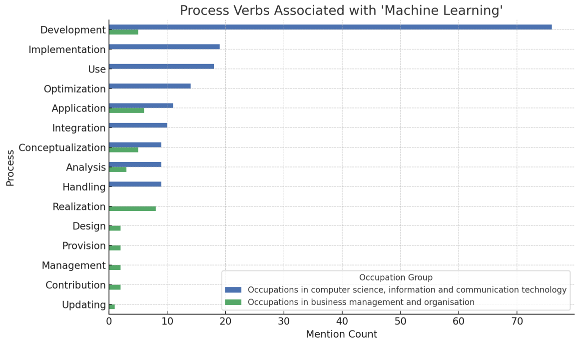

Figure 5: Process verbs associated with *machine learning* across occupational domains.

ing is embedded in different occupational domains. We compare two groups based on the German classification of occupations (KldB): *Occupations in business management and organisation* and *Occupations in computer science, information and communication technology*.

Figure 5 shows the process verbs most frequently associated with *machine learning* in both groups, aggregated across lexical variants. In ICT-related occupations, machine learning is predominantly linked to development-oriented processes such as *developing*, *implementing*, and *optimizing*. In contrast, business-related roles emphasize more strategic or organisational actions such as *realizing*, *applying*, or *conceptualizing*.

PyTorch *or* TensorFlow? Our schema captures logical relations such as disjunctions, e.g., in phrases like "experience with PyTorch *or* TensorFlow".

Among job ads mentioning both frameworks, 57.4% explicitly encode this as an `Alternative` relation—indicating that only one is required. The remaining 42.6% list both without a linking relation, leaving the requirement ambiguous.

How urgent is AI? To assess the framing of AI-related experience requirements, we analyzed annotation chains of the form:

$\texttt{Work Content} \rightarrow \texttt{Experience Level} \xrightarrow{\texttt{Urgency}} \texttt{Attribute}$

For each `Attribute`, we applied a zero-shot classification using an mDeBERTa-based NLI model (Laurer et al., 2024). Based on the surface form of the attribute (e.g., "nice to have", "required", "ideally"), we assigned one of three urgency levels: *required*, *preferable*, or *unimportant*.

Table 3 shows that only 1.4% of AI-related cases are marked as *required*, while 97.2% are *preferable*.

Type	Required	Unimportant	Preferable
AI-related	1.4%	1.4%	97.2%
Non-AI-related	8.9%	2.5%	88.6%

Table 3: Distribution of urgency classifications for experience-related requirements based on NLI predictions over structured entity chains.

Non-AI mentions more often indicate mandatory expectations.

These findings suggest that AI is still largely framed as an optional asset, reflecting early-stage adoption. This helps explain how emerging technologies enter occupational profiles—first as desirable attributes, later as standardized requirements.

Summary These examples demonstrate the analytical value of our schema and dataset, enabling the exploration of semantically rich questions. Analyses like modeling *urgency* or identifying *alternatives* are only accessible through structured annotations. While single-span approaches might approximate them via inference pipelines (cf. Section 4), our schema captures such distinctions natively and directly.

We acknowledge that the first two examples, involving *robotics* and *machine learning*, could in principle also be distinguished through normalized flat outputs, even though we did not perform taxonomy normalization in this study. Nevertheless, the structural clarity of our schema simplifies such normalization and facilitates direct integration into taxonomies—particularly in the presence of long spans, conjoined expressions, or ambiguous structures (cf. Section 1, 2.4).

Beyond facilitating analysis, the structured output also supports taxonomy development itself: by applying these methods to larger datasets and clustering co-occurring process expressions, empirical structures can inform or revise existing classification systems. Finally, we emphasize that this study is a proof of concept. Several entity types, such as `Formal Qualification`, `Job Title`, or `Sector`, as well as longer relational chains, remain unexplored—highlighting the substantial potential for future work.

4 Discussion

Our findings confirm that compositional modeling is not only conceptually well-founded but also empirically feasible and analytically valuable. GOJA demonstrates that detailed, structured representations of requirements can be annotated with high reliability and effectively predicted by transformer models. It should be noted, however, that comparability with previous work—such as Zhang et al. (2023)—is limited, as most existing approaches rely on flat span-based annotation of isolated concepts. Reported extraction performance in these studies varies widely depending on how skills are defined, with simpler formulations often yielding higher scores at the cost of structural and semantic depth (cf. Alexopoulos, 2020).

Emerging compositional approaches in OJA research. Recent studies have begun to address the structural limitations of single-span extraction. As shown in Figure 1 Zhang et al. (2022b) extend span-based labeling by allowing nested annotations, while Nguyen et al. (2024) formulate extraction as a generative task to improve flexibility. Gnehm et al. (2022a) demonstrate that deeper semantic patterns can indeed be extracted from flat annotations — but only through additional decomposition steps that segment and classify subcomponents of long spans post hoc. Compared to these approaches, our method offers several concrete advantages: it is more efficient than generative models, as it relies on standard encoder-based architectures; it handles conjoined expressions more reliably (Nguyen et al., 2024) by representing them structurally; and it enables selective modeling of relevant information—allowing the model to ignore contextually unimportant modifiers. Crucially, our schema encodes explicit semantic relationships, which not only increases representational richness and accurateness but also supports new types of research questions, as demonstrated in our case study on AI-related requirements.

Broader applicability. Compositional modeling of entities and concepts is not unique to our approach; it also underlies many relation extraction tasks where relations between entities construct higher-order concepts. While traditional relation extraction typically operates on classic named entities, our method starts from predefined conceptual structures and decomposes them into text-based components. Despite differences in granularity, both approaches transform lower-level units into more complex representations.

Unlike traditional relation extraction approaches that usually operate over classic named entities, our method starts from conceptual units and builds interpretable structures over text spans.We believe that the broader NLP community, particularly in application-driven fields such as industry, computational social science (CSS), and digital humanities (DH), could benefit from a more extensive discussion on compositionality in text and its relation to conceptual modeling. Our findings highlight the limitations of treating many information extraction IE tasks purely as named entity recognition (NER) problems.

5 Conclusion and Outlook

This paper introduced a compositional entity modeling framework for requirement extraction from Online Job Advertisements (OJAs). Rather than modeling requirements as isolated spans, our approach captures their internal structure by annotating typed entities and their semantic relations. Based on this framework, we present GOJA, a gold-standard dataset of 500 annotated German job ads, demonstrating high annotation consistency and the feasibility of training extraction models on this structured representation.

Our work opens several avenues for future research. While our dataset focuses on German OJAs, future studies could explore whether compositional modeling yields similar benefits across languages and domains. More extensive benchmarking, including additional evaluation metrics (e.g., triple-level accuracy), aggregation of higher-order concepts (e.g., tasks and skills), and advanced architectures (e.g., joint entity-relation extraction or graph-based models Shaowei et al., 2022; Wu et al., 2020), could provide further insights.

Beyond extraction, requirement modeling often involves aligning extracted content with external taxonomies or ontologies. Since such resources can be represented as graphs (see Dörpinghaus et al., 2023), the structured output of our schema — including relational chains and alternatives — may support hierarchical or joint taxonomy alignment. Furthermore, our case study already illustrated the analytical potential of structured representations; scaling this approach to larger and longitudinal datasets may enable systematic investigations into emerging skills, requirement trends, and taxonomy alignment.

In conclusion, our framework contributes a robust foundation for analyzing complex requirements in job advertisements and encourages broader discussion around compositional representations in applied information extraction tasks.

6 Limitations

While our compositional entity modeling framework shows promising results in capturing complex semantic dependencies in online job advertisements, several limitations and deliberate design decisions should be acknowledged.

Limited Large-Scale Empirical Validation. Although our experiments indicate that the proposed method can more effectively capture the intricate structure of job requirements compared to flat entity extraction methods, conclusively validating this claim would require large-scale empirical comparisons across diverse modeling paradigms. Such an endeavor would involve developing and benchmarking multiple models on datasets comprising millions of OJAs and assessing their performance across various downstream applications (e.g., skill gap analysis, regional labor market assessments). Given the substantial scope and resource requirements, this comprehensive evaluation remains beyond the scope of the current study.

Design Decisions in Entity and Relation Definitions. A central design choice of our framework is to consistently label similar textual components with the same entity type—specifically, using work content for elements that denote the object or subject within a sentence. For example, a machine mentioned in a job advertisement is always annotated as Work content, irrespective of whether the context involves repairing or operating machinery. The semantic differences between these contexts are then captured through distinct relation types: when the machine is directly acted upon (as in repairing machinery), the relation OBT is used, whereas if it serves as an instrument (as in operating machinery), the relation Tool is applied. This choice was made, because we believe it would enhance annotation consistency and model performance.

Then, other relational distinctions, such as Alternative, emerge directly from the logical structure of the text. However, decisions regarding when to introduce a new entity versus representing semantic nuances solely through relations (e.g., the case of Specialization, which often maps to attributes) proved challenging and, in some cases, inherently arbitrary. These design choices could affect both the generalizability of the framework and the interpretability of the extracted structures. Balancing the need for annotation consistency with the capture of fine-grained semantic distinctions remains an open challenge and a potential limitation of our approach.

Context Window and Sentence Splitting. For relation classification, we sample candidate entity pairs within a context window defined by sentence boundaries. This decision was based on analyses suggesting that sentences provide a natural and less arbitrary segmentation unit compared to tokens or words. However, sentence splitting in job advertisements is challenging due to unconventional punctuation, enumerations, and gender-neutral formulations in German. Such issues can lead to suboptimal context sizes, potentially affecting the capture of relevant relational dependencies. Future work should investigate more robust segmentation strategies.

Token Alignment Issues. Our annotations are performed at the character level and subsequently aligned with tokenized text. In rare cases, discrepancies between token boundaries and annotated spans occur. Although internal analysis indicates that these misalignments are marginal, they nonetheless represent a potential source of error that might slightly affect extraction performance during inference. Addressing these alignment challenges is an important direction for future research. Note, that this problem did not affect the model performances presented in Section 3.1.

Comparison with Flat Entity Extraction. A potential counterargument is that extracting longer spans as single units might allow for semantic and logical connections to be resolved in downstream processing. However, research (Zhang et al., 2022b) has shown that longer, compositionally rich spans are increasingly difficult for models to extract reliably. Thus, while flat entity extraction may delay the need to capture internal structure, it does not remove the underlying challenge of representing complex requirement semantics in job advertisements.

References

Panos Alexopoulos. 2020. *Semantic Modeling for Data*. O'Reilly Media.

Enghin Atalay, Phai Phongthiengtham, Sebastian Sotelo, and Daniel Tannenbaum. 2018. New technologies and the labor market. *Journal of Monetary Economics*, 97:48–67.

Enghin Atalay, Phai Phongthiengtham, Sebastian Sotelo, and Daniel Tannenbaum. 2020. The evolution of work in the united states. *American Economic Journal: Applied Economics*, 12(2):1–34.

Enghin Atalay, Sebastian Sotelo, and Daniel Tannenbaum. 2023. The geography of job tasks. *Journal of Labor Economics*.

David H Autor and Michael J Handel. 2013. Putting tasks to the test: Human capital, job tasks, and wages. *Journal of labor Economics*, 31(S1):S59–S96.

Rahkakavee Baskaran and Johannes Müller. 2023. Classification of german job titles in online job postings using the kldb2010 taxonomy. Last accessed: 2024-05-22.

David Beauchemin, Julien Laumonier, Yvan Le Ster, and Marouane Yassine. 2022. " fijo": a french insurance soft skill detection dataset. *arXiv preprint arXiv:2204.05208*.

Adam SZ Belloum, Spiros Koulouzis, Tomasz Wiktorski, and Andrea Manieri. 2019. Bridging the demand and the offer in data science. *Concurrency and Computation: Practice and Experience*, 31(17):e5200.

Akshay Bhola, Kishaloy Halder, Animesh Prasad, and Min-Yen Kan. 2020. Retrieving skills from job descriptions: A language model based extreme multi-label classification framework. In *Proceedings of the 28th international conference on computational linguistics*, pages 5832–5842.

Katy Börner, Olga Scrivner, Mike Gallant, Shutian Ma, Xiaozhong Liu, Keith Chewning, Lingfei Wu, and James A. Evans. 2018. Skill discrepancies between research, education, and jobs reveal the critical need to supply soft skills for the data economy. *Proceedings of the National Academy of Sciences of the United States of America*, 115(50):12630–12637.

Phillip Brown and Manuel Souto-Otero. 2020. The end of the credential society? an analysis of the relationship between education and the labour market using big data. *Journal of Education Policy*, 35(1):95–118.

Marlis Buchmann, Helen Buchs, Felix Busch, Simon Clematide, Ann-Sophie Gnehm, and Jan Müller. 2022. Swiss job market monitor: A rich source of demand-side micro data of the labour market. *European Sociological Review*, 38(6):1001–1014.

Emilio J Castilla and Hye Jin Rho. 2023. The gendering of job postings in the online recruitment process. *Management Science*, 69(11):6912–6939.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jens-Joris Decorte, Jeroen Van Hautte, Johannes Deleu, Chris Develder, and Thomas Demeester. 2022. Design of negative sampling strategies for distantly supervised skill extraction. *arXiv preprint arXiv:2209.05987*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jens Dörpinghaus, Johanna Binnewitt, Stefan Winnige, Kristine Hein, and Kai Krüger. 2023. Towards a german labor market ontology: Challenges and applications. *Applied Ontology*, (Preprint):1–23.

Marta Fana, Martina Bisello, Sergio Torrejón Pérez, and Enrique Fernández-Macías. 2023. What workers do and how. a european database of tasks indices.

Anna Giabelli, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Andrea Seveso. 2021. Neo: A system for identifying new emerging occupation from job ads. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(18):16035–16037.

Ann-Sophie Gnehm. 2018. Text zoning for job advertisements with bidirectional lstms. In *Proceedings of the 3rd Swiss Text Analytics Conference (SwissText 2018)*, pages 1–9, Winterthur. University of Zurich.

Ann-sophie Gnehm, Eva Bühlmann, Helen Buchs, and Simon Clematide. 2022a. Fine-grained extraction and classification of skill requirements in german-speaking job ads. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+ CSS)*. Association for Computational Linguistics.

Ann-Sophie Gnehm, Eva Bühlmann, and Simon Clematide. 2022b. Evaluation of transfer learning and domain adaptation for analyzing german-speaking job advertisements. In *Proceedings of the 13th Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association.

Ann-Sophie Gnehm and Simon Clematide. 2020. Text zoning and classification for job advertisements in german, french and english. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 83–93.

Thomas AF Green, Diana Maynard, and Chenghua Lin. 2022. Development of a benchmark corpus to support entity recognition in job descriptions. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1201–1208. European Language Resources Association.

Betül Güntürk-Kuhl, Philipp Martin, and Anna Cristin Lewalder. Die taxonomie der arbeitsmittel des bibb: Revision 2018.

Imane Khaouja, Ismail Kassou, and Mounir Ghogho. 2021. A survey on skill identification from online job ads. *IEEE Access*, 9:118134–118153.

Jeonghyun Kim and Putthachat Angnakoon. 2016. Research using job advertisements: A methodological assessment. *Library & Information Science Research*, 38(4):327–335.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Moritz Laurer, Wouter Van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100.

Nan Li, Bo Kang, and Tijl de Bie. Llm4jobs: Unsupervised occupation extraction and standardization leveraging large language models.

Antonio Lima, B Bakhshi, et al. 2018. Classifying occupations using web-based job advertisements: an application to stem and creative occupations. *Economic Statistics Centre of Excellence Discussion Paper*, 8.

Felix Lukowski, Myriam Baum, and Sabine Mohr. 2021. Technology, tasks and training–evidence on the provision of employer-provided training in times of technological change in germany. *Studies in Continuing Education*, 43(2):174–195.

Johannes Müller. Machbarkeitsstudie: Teilqualifikationen in online-job-anzeigen (oja): Methodenbericht zur automatisierten extraktion von teilqualifikationen für fünf ausbildungsberufe: Projekt: Aufstieg durch kompetenzen.

Khanh Nguyen, Mike Zhang, Syrielle Montariol, and Antoine Bosselut. 2024. Rethinking Skill Extraction in the Job Market Domain using Large Language Models. In *Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024)*, pages 27–42, St. Julian's, Malta. Association for Computational Linguistics.

Dimos Ntioudis, Panagiota Masa, Anastasios Karakostas, Georgios Meditskos, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2022. Ontology-based personalized job recommendation framework for migrants and refugees. *Big Data and Cognitive Computing*, 6(4):120.

Lennert Peede and Michael Stops. 2024. Artificial intelligence technologies, skills demand and employment: evidence from linked job ads data. Technical report, IAB-Discussion Paper.

Ibrahim Rahhal, Kathleen M. Carley, Ismail Kassou, and Mounir Ghogho. 2023. Two stage job title identification system for online job advertisements. *IEEE Access*, 11:19073–19092.

Nils Reiter, Marcus Willand, and Evelyn Gius. 2019. A shared task for the digital humanities chapter 1: Introduction to annotation, narrative levels and shared tasks. *Journal of Cultural Analytics*, 4(3).

Margarida Rodrigues, Fernández-Macías, and Enrique, Sostero, Matteo. 2021. A unified conceptual framework of tasks, skills and competences.

Angelo A Salatino, Thiviyan Thanapalasingam, Andrea Mannocci, Francesco Osborne, and Enrico Motta. 2018. The computer science ontology: a large-scale taxonomy of research areas. In *The Semantic Web–ISWC 2018: 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part II 17*, pages 187–205. Springer.

V Sanh. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Luiza Sayfullina, Eric Malmi, and Juho Kannala. 2018. Learning representations for soft skill matching. In *Analysis of Images, Social Networks and Texts: 7th International Conference, AIST 2018, Moscow, Russia, July 5–7, 2018, Revised Selected Papers 7*, pages 141–152. Springer.

Benjamin Schimke. 2023. Nachweise für berufliche qualifikationen oder doch nur ein motivationssignal? zur wirkung non-formaler weiterbildungszertifikate in der personalauswahl. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 75(4):451–475.

Elena Senger, Mike Zhang, Rob van der Goot, and Barbara Plank. 2024. Deep learning-based computational job market analysis: A survey on skill extraction and classification from job postings. *Preprint*, arXiv:2402.05617.

ZHANG Shaowei, WANG Xin, CHEN Zirui, WANG Lin, XU Dawei, and JIA Yongzhe. 2022. Survey of supervised joint entity relation extraction methods. *Journal of Frontiers of Computer Science & Technology*, 16(4).

Ellery Smith, Andreas Weiler, and Martin Braschler. 2021. Skill extraction for domain-specific text retrieval in a job-matching platform. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12*, pages 116–128. Springer.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24.

Mike Zhang, Kristian Nørgaard Jensen, and Barbara Plank. 2022a. Kompetencer: Fine-grained skill classification in danish job postings via distant supervision and transfer learning. In *13th International Conference on Language Resources and Evaluation*, pages 436–447. European Language Resources Association (ELRA).

Mike Zhang, Kristian Nørgaard Jensen, Sif Dam Sonniks, and Barbara Plank. 2022b. Skillspan: Hard and soft skill extraction from english job postings. In *2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Mike Zhang, Rob van der Goot, and Barbara Plank. 2023. ESCOXLM-R: Multilingual taxonomy-driven pre-training for the job market domain. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11871–11890, Toronto, Canada. Association for Computational Linguistics.

A Dataset Details

Data Sampling. To reduce biases, for example due to data shift or OJAs differing between jobs or industry sectors, we applied a multivariate sampling approach. Table 4 explains the different variables used.

Analysis of Conjoined Structures To illustrate the structural complexity of requirement expressions in Online Job Advertisements (OJAs), Figure 6 presents a breakdown of frequently observed conjoined patterns. These include, for example, single processes linked to multiple work contents, or experience levels associated with multiple tasks or tools. The visualization aggregates entity chains into abstracted patterns to support interpretability.

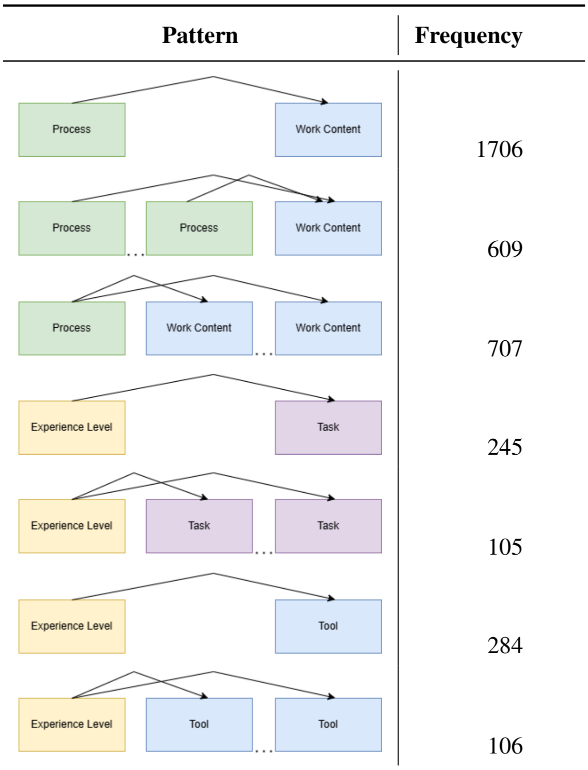

Figure 6: Frequency of conjoined requirement structures in GOJA. Each pattern groups structurally similar chains; entities denoted as n represent arbitrarily many nodes of the same type.

Annotation guidelines. Annotation guidelines can be accessed under `https://github.com/TM4VETR/Public_Stea_Annotationsguide`

Annotators. All annotators (A+B) work in the same organization as the authors of this article. They are all native German speakers and hold at least the equivalent of a Bachelor's degree, with diverse backgrounds in social sciences, (digital) humanities, economics, and psychology. All have at least some experience in labor market research, which is advantageous given the complex structure of the operationalization of the concepts. Four of the annotators are male, and eleven are female.

All annotations were conducted during regular working hours, and the annotators did not receive any additional payment beyond their regular salary. All B participated voluntarily following a call for participation.

The annotators were informed about the purpose of the annotation process, and in exchange for their contribution, they were promised priority access to the final dataset.

Additional IAA scores. Tables 5 and 6 show the IAA results per class.

Entity and relation counts. Table 7 displays of the amount of annotated entities and relations in our dataset.

B Experimental Setup Details

To ensure reproducibility, we provide additional details on our experimental setup:

Hyperparameters. Table 8 and Table 9 provide details regarding the hyperparameters used in our experiments.

Hardware: All models were trained on an NVIDIA L40 GPU with 48 GB VRAM.

Class Imbalance: The "No Relation" class was downsampled to match the total number of instances in other relation classes.

Cross-Validation: A stratified 5-fold cross-validation was performed using the same five random seeds across all models.

Licences:

C Additional Analysis

Figures 7 and 8 display the aggregated confusion matrices for entity extraction and relation classification, respectively, across five runs per model. As they do use numeric labels for space reasons, the label mapping presented in Tables 11 and 12 respectively.

C.0.1 Error Analysis

Our error analysis aims to explain model performance differences on a per-class level and to understand the relationship between model predictions, inter-annotator agreement (IAA), and error patterns. Figure 9 presents per-class F1 scores and std. deviations, while confusion matrices (Figures 7 and 8) illustrate detailed prediction errors. Our

Factor	Description
Year of Publishing	Job ads from the years 2016 and 2022.
Source Website	Job portals and company websites.
WZ08 Activity	Selection from the economic sections of the WZ08 classification.
ISCO08 Occupation	First level of the ISCO08 occupational classification.
Contract Type	Only permanent and fixed-term contracts (excluding apprenticeships, internships, etc.).
Text Length	Various text lengths, measured using spaCy tokenization.

Table 4: Factors in the Multivariate Sampling Approach for Job Ad Selection

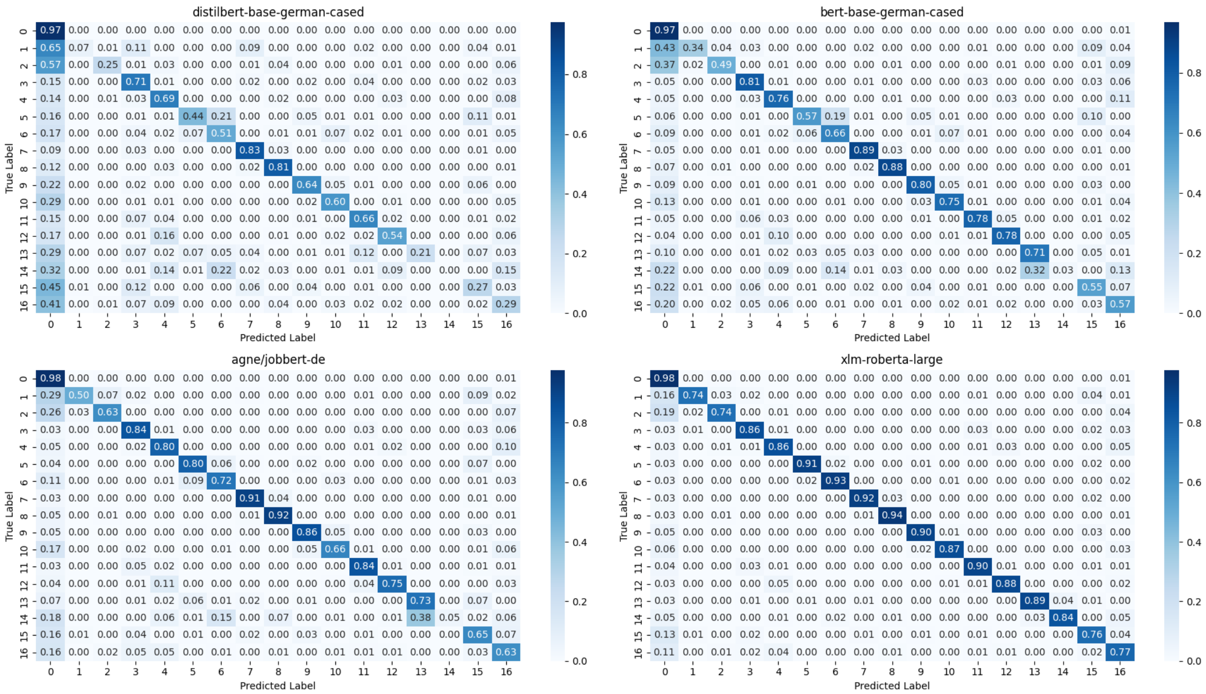

Figure 7: Aggregated confusion matrices for entity extraction (row-normalized over 5 runs for each model)

Entity Type	IAA (Krippendorff's α)
Work Content	0.75
Attitude	0.87
Attribute	0.60
Occupation	0.83
Industry	0.55
Experience Level	0.85
Formal Qualification	0.87
Process	0.78

Table 5: Inter-Annotator Agreement (Krippendorff's α) for Entity Types

analysis shows that superior macro-F1 scores of XLM-RoBERTa stem primarily from its ability to handle difficult classes rather than from general peak performance.

Weak classes. Entity extraction errors cluster around three difficult classes: Formal Qualification (FQ), Attribute, and Industry. Relation extraction errors are concentrated in Degree of Autonomy and REP. Attribute and Industry are conceptually difficult, reflected in low IAA scores. Attribute acts as a broad, catch-all category with long and inconsistent spans, while Industry annotations are limited to candidate-focused sections, causing ambiguity about what qualifies as an industry mention. Both classes are frequently confused with the Outside (O) label, as shown in the confusion matrices, which is less critical since these errors often reflect borderline cases rather than clear misclassifications.

A similar pattern appears in relation classification: Degree of Autonomy and REP have low IAA scores and few examples, resulting in low F1 scores.

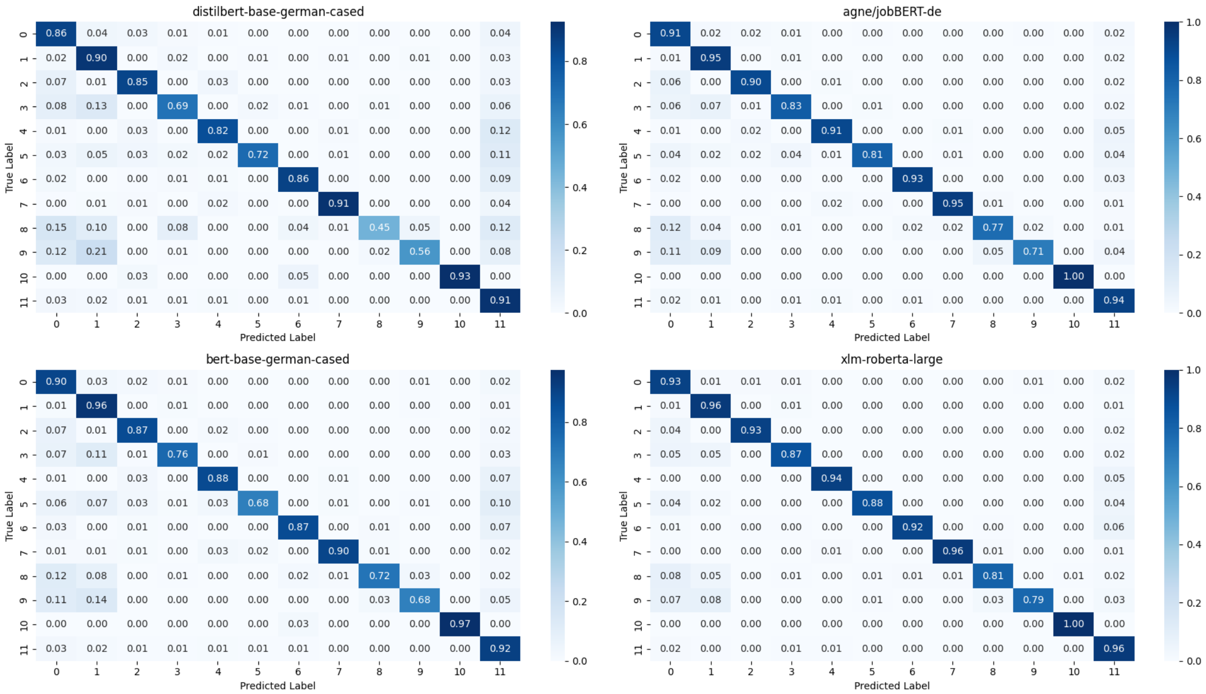

Figure 8: Aggregated confusion matrices for relation classification (row-normalized over 5 runs for each model)

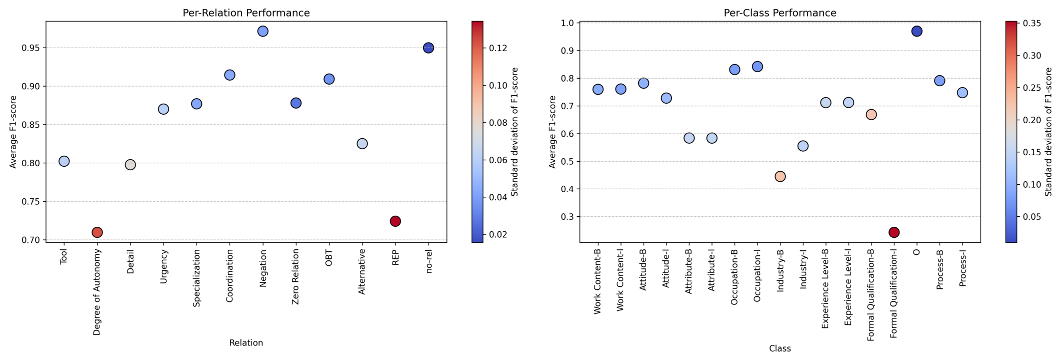

Figure 9: Mean F1-score across all models for each entity and relation class. The color gradient represents the standard deviation of F1-scores across runs.

In contrast, other classes with low IAA scores, such as Zero Relation and Specialization, perform better due to having more examples. The high performance of Negation, despite having few examples, further suggests that performance depends on both conceptual clarity and class frequency.

FQ as a notable outlier. Although the FQ class exhibits high IAA scores and clear conceptual boundaries, it performs poorly for all models except XLM-RoBERTa. Confusion matrices reveal that weaker models seldom predict `FQ-I` at all. Besides the general overprediction of the outside class, the models show different behavior in regard to FQ. DistilBERT models frequently predict Work Content-I, Attribute-I, or Experience Level-I instead of FQ-I. Manual inspection shows that these models often switch from FQ-B to the inside tag of another entity type mid-span. Both the internal splitting of spans and the confusion between semantically distinct entity types are notable and unexpected. In contrast, BERT and jobBERT-de models display a different error pattern: they tend to predict FQ-B but fail to continue the span with FQ-I, predicting another FQ-B. Only XLM-RoBERTa is able to predict FQ reliably.

D Information About Use Of AI Assistants

We used AI assistants as a tool to support both the writing and coding aspects of this research. In particular, AI-assisted tools were employed to gen-

Relation Type	**IAA (Krippendorff's α)**
Alternative	0.75
Coordination	0.75
Degree of Autonomy	0.62
Detail	0.62
Negation	0.90
Object Being Transformed (OBT)	0.72
Related Entity Parts (REP)	0.67
Specialization	0.68
Tool	0.61
Urgency	0.78
Zero Relation	0.52

Table 6: Inter-Annotator Agreement (Krippendorff's α) for Relation Types

erate initial drafts of text, suggest improvements in language and structure, and assist with coding tasks. All AI-generated content was thoroughly reviewed, refined, and integrated by the authors to ensure accuracy, clarity, and alignment with our research objectives. The use of AI was solely aimed at increasing efficiency in routine tasks, and final decisions and edits were made by the research team.

E Ethics statement

Our study is purely academic in nature, and we do not foresee any significant risks or adverse impacts arising from our approach. The dataset used consists of non-public job advertisements and has been processed strictly for research purposes, with all sensitive information anonymized prior to analysis. Given that our methodology is applied solely for analytical and evaluation objectives, we believe that our work does not pose any harm.

Entities	**Count**
Work Content	5285
Attribute	4685
Process	4461
Attitude	2172
Occupation	2105
Industry	1615
Experience Level	1412
Formal Qualification	771
Relations	**Count**
Zero Relation	4322
OBT	3648
Specialization	1345
Tool	1157
Alternative	597
Detail	585
Coordination	482
Urgency	466
Degree of Autonomy	325
REP	312
Negation	85

Table 7: Number of annotated entities and relations per class

Task	**XLM-RoBERTa**	**jobBERT-de, German BERT**	**DistilBERT**
Entity Extraction	7 epochs	9 epochs	15 epochs
Relation Classification	6 epochs	8 epochs	12 epochs

Table 8: Number of epochs per model

Hyperparameter	**Value**
Batch Size	64 (XLM-RoBERTa: 16)
Learning Rate	5e-5
Weight Decay	0
Adam Betas	(0.9, 0.999)
Adam Epsilon	1e-8
Max Gradient Norm	1.0
Scheduler	Linear
Warmup Ratio	0.0

Table 9: Hyperparameter details

Model	**License**
MoritzLaurer/DeBERTa-v3-base-mnli-fever-docnli-ling-2c	MIT License
google-bert/bert-base-german-cased	MIT License
distilbert/distilbert-base-german-cased	Apache License 2.0
agne/jobBERT-de	CC-BY-NC-SA 4.0
FacebookAI/xlm-roberta-base	MIT License

Table 10: Model licences

Label number	**Label name**
0	O
1	Industry-B
2	Industry-I
3	Work Content-B
4	Work Content-I
5	Experience Level-B
6	Experience Level-I
7	Occupation-B
8	Occupation-I
9	Attitude-B
10	Attitude-I
11	Process-B
12	Process-I
13	Formal Qualification-B
14	Formal Qualification-I
15	Attribute-B
16	Attribute-I

Table 11: Entity label mapping

Label number	**Label number**
0	Zero Relation
1	OBT
2	Specialization
3	Tool
4	Alternative
5	Detail
6	Urgency
7	Coordination
8	REP
9	Degree of Autonomy
10	Negation
11	no-rel

Table 12: Relation label mapping