**Introduction & Motivation**

- Skill Extraction (SE) is the task of extracting spans from job ads. Certain skills might be underrepresented in job description, resulting in a sparsity of skills SE datasets.
- In job descriptions, there is a long-tail pattern, popular skills are more commonly mentioned, while niche expertise appears less frequently.
- We explore Nearest Neighbor Language Models (NNLMs; Khandelwal et al., 2020), using the kNN algorithm as a retriever to retrieve context—token pairs from a datastore with LM encoders.

**Models & Data**

- JobBERT (Zhang et al., 2022)
- RoBERTa (Liu et al., 2019)
- JobBERTa (This work): RoBERTa further pre-trained on 3.2M job posting sentences.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>D (tokens)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SkillSpan</td>
<td>5,866</td>
<td>3,992</td>
<td>4,680</td>
<td>86.5K</td>
</tr>
<tr>
<td>Sayfullina</td>
<td>3,706</td>
<td>1,854</td>
<td>1,853</td>
<td>53.1K</td>
</tr>
<tr>
<td>Green</td>
<td>8,670</td>
<td>963</td>
<td>336</td>
<td>209.5K</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td>348.2K</td>
</tr>
</tbody>
</table>

Table 1: Dataset Statistics. We provide statistics for all three datasets. Input granularity is at the token level, with performance measured in span-F1. The size of the datastore $D$ is in tokens and determined by embedding tokens and their context from the training sets, resulting in approximately 350K keys.

**Approach**

- We observe large gains using NNOSE in a cross-dataset setting.
- We train on one dataset and apply the model to another using the datastore.
- We confirm findings similar to Khandelwal et al. (2020), that memorisation using NNLMs improves recall.