cases	doc_1	doc_2	decision	id
	<table><tr><td>**authors**</td><td>• KarolÃna BeneÅ¡ovÃ¡ • Andrej Å vec • Marek Å uppa</td></tr><tr><td>**title**</td><td>Cost-effective Deployment of BERT Models in Serverless Environment</td></tr><tr><td>**publication_date**</td><td>2021-03-19 00:00:00</td></tr><tr><td>**source**</td><td>SupportedSources.OPENALEX</td></tr><tr><td>**journal**</td><td>arXiv (Cornell University)</td></tr><tr><td>**volume**</td><td></td></tr><tr><td>**doi**</td><td>None</td></tr><tr><td>**urls**</td><td>• https://openalex.org/W3137254616</td></tr><tr><td>**id**</td><td>id-1879542727168961955</td></tr><tr><td>**abstract**</td><td></td></tr><tr><td>**versions**</td><td></td></tr></table>	<table><tr><td>**authors**</td><td>• KatarÃna BeneÅ¡ovÃ¡ • Andrej Å vec • Marek Å uppa</td></tr><tr><td>**title**</td><td>Cost-effective Deployment of BERT Models in Serverless Environment</td></tr><tr><td>**publication_date**</td><td>2021-03-19 07:45:17+00:00</td></tr><tr><td>**source**</td><td>SupportedSources.ARXIV</td></tr><tr><td>**journal**</td><td>None</td></tr><tr><td>**volume**</td><td></td></tr><tr><td>**doi**</td><td></td></tr><tr><td>**urls**</td><td>• http://arxiv.org/pdf/2103.10673v2 • http://arxiv.org/abs/2103.10673v2 • http://arxiv.org/pdf/2103.10673v2</td></tr><tr><td>**id**</td><td>id4327090977737325995</td></tr><tr><td>**abstract**</td><td>In this study we demonstrate the viability of deploying BERT-style models to serverless environments in a production setting. Since the freely available pre-trained models are too large to be deployed in this way, we utilize knowledge distillation and fine-tune the models on proprietary datasets for two real-world tasks: sentiment analysis and semantic textual similarity. As a result, we obtain models that are tuned for a specific domain and deployable in serverless environments. The subsequent performance analysis shows that this solution results in latency levels acceptable for production use and that it is also a cost-effective approach for small-to-medium size deployments of BERT models, all without any infrastructure overhead.</td></tr><tr><td>**versions**</td><td></td></tr></table>	DUPLICATES	188