pjmathematician at TalentCLEF 2025: Enhancing Job Title and Skill Matching with GISTEmbed and LLM-Augmented Data

Notebook for the TalentCLEF Lab 2025

Poojan Vachharajani[1]

[1]*Netaji Subhas University of Technology, New Delhi, India*

Abstract

This paper details the pjmathematician team's participation in the TalentCLEF 2025 shared task, focusing on Task A (Multilingual Job Title Matching) and Task B (Job Title-Based Skill Prediction). Our approach primarily leveraged state-of-the-art sentence embedding models fine-tuned using the GISTEmbed technique. For Task A, various multilingual and English-specific encoder models were adapted, including a distilled version and a LoRA-fine-tuned 7B parameter model. Data augmentation for Chinese was performed using Qwen2.5 32B Instruct. For Task B, we employed data augmentation using Qwen2.5 32B Instruct to generate descriptive texts for jobs and skills, significantly enriching the training data. Models like BGE-Large and GTE-Qwen2-7B (LoRA) were fine-tuned on this augmented data. Our submissions demonstrate the effectiveness of these strategies, achieving competitive results in both tasks.

Keywords

TalentCLEF, Job Title Matching, Skill Prediction, Sentence Embeddings, GISTEmbed, Data Augmentation, Large Language Models, LoRA

1. Introduction

The TalentCLEF 2025 shared task [1] presents challenges in leveraging AI for human capital management. This paper describes the participation of the pjmathematician team in two sub-tasks: Task A, Multilingual Job Title Matching, and Task B, Job Title-Based Skill Prediction. Task A requires systems to identify and rank job titles similar to a given query across English, Spanish, German, and Chinese. Task B focuses on retrieving relevant skills for a given job title in English.

Our general approach involves fine-tuning various pre-trained sentence embedding models [2] using the GISTEmbed [3] loss function. For Task A, we explored several base models including BGE [4, 5] and GTE [6, 7] series, employing techniques like LoRA fine-tuning for larger models and knowledge distillation for model compression. External data from ESCO was used, particularly for generating Chinese training data via LLM-based translation. For Task B, a significant component of our strategy was data augmentation using the Qwen2.5 32B Instruct model to generate rich descriptions for jobs and skills, thereby enhancing the training dataset for fine-tuning models like BGE-Large and GTE-Qwen2-7B. All models were implemented using the sentence-transformers library [2].

2. Methodology

2.1. Task A: Multilingual Job Title Matching

Our approach for Task A centered on fine-tuning various sentence embedding models to capture semantic similarities between job titles across multiple languages.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

pjmathematician@gmail.com (P. Vachharajani)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

CEUR-WS.org/Vol-4038/paper_375.pdf

2.1.1. Models and Fine-tuning

We experimented with several base encoder models:

- BAAI/bge-small-en-v1.5 (33.4M parameters) [4]
- BAAI/bge-m3 (569M parameters) [5]
- Alibaba-NLP/gte-multilingual-base (305M parameters) [6]
- Alibaba-NLP/gte-Qwen2-7B-instruct (7B parameters) [7]
- A distilled version of Alibaba-NLP/gte-multilingual-base (approx. 60M parameters).

All models, except the distilled one, were fine-tuned using the GISTEmbed loss [3], with ‘all-MiniLM-L12-v2‘ [2] as the guide model. For the GTE-Qwen2-7B-instruct model, CachedGISTEmbed was used along with Low-Rank Adaptation (LoRA) [8] to manage computational resources. The distilled ‘gte-multilingual-base‘ model (retaining 3 layers) was fine-tuned using Mean Squared Error (MSE) loss to mimic the embeddings of its GISTEmbed-fine-tuned, full-layer counterpart which served as the teacher model.

2.1.2. Data

The provided training set of related job title pairs for English, Spanish, and German was used. For Chinese, where no training data was provided, we augmented the ESCO English job title dataset by translating it to Chinese using the Qwen2.5 32B Instruct model. All language pairs were used to train a single multilingual model for each respective base encoder.

2.1.3. Implementation

All fine-tuning and inference were performed using the ‘sentence-transformers‘ library [2]. Bias handling involved shuffling the training data and using random validation splits.

2.2. Task B: Job Title-Based Skill Prediction

For Task B, our strategy focused on robust data augmentation using a Large Language Model (LLM) and fine-tuning powerful English-specific sentence encoders.

2.2.1. Data Augmentation

We utilized the Qwen2.5 32B Instruct (AWQ quantized) model to generate descriptive text for both job titles and skills. The process involved prompting the LLM to create concise descriptions:

- For jobs: "Given a job role (and its synonyms), briefly (1-2 sentences) describe the skills needed for that job". This generated a ‘skill_brief‘ for each job title.
- For skills: "Given a skill (and its synonyms), briefly (1-2 sentences) describe the job roles that require that skill". This generated a ‘job_brief‘ for each skill.

This process was applied to the training, validation, and test sets provided by the organizers. Two main augmented training datasets were created from the original job-skill mapping files:

1. A dataset of pairs (‘skill_brief‘, ‘job_brief‘).
2. A dataset of mixed-format pairs (‘skill_brief‘ + newline + job synonyms list, ‘job_brief‘ + newline + skill synonyms list).

2.2.2. Models and Fine-tuning

We employed two main base models:

- **BAAI/bge-large-en-v1.5 (335M parameters) [4]:**
  - One version was fine-tuned using CachedGISTEmbedLoss on the dataset composed of augmented skill and job descriptions. Inference used these generated descriptions.
  - Another version was fine-tuned using CachedGISTEmbedLoss on the dataset of mixed augmented descriptions and synonym lists. Inference used inputs combining the generated description with the original job title or skill aliases.
- **Alibaba-NLP/gte-Qwen2-7B-instruct (7B parameters) [7]:** This model was fine-tuned using LoRA and CachedGISTEmbedLoss on the dataset of mixed augmented descriptions and synonym lists. Inference input combined the generated description with the original job title or skill aliases.

For all GISTEmbed fine-tuning, 'all-MiniLM-L12-v2' [2] served as the guide model.

2.2.3. Implementation and Inference

The 'sentence-transformers' library [2] was used for training and inference. Submissions were generated by encoding the augmented query and corpus texts and then computing cosine similarity scores to rank corpus elements. No external data beyond the LLM-generated augmentations was used for Task B. Shuffling and random validation splits were standard practice.

3. Experiments and Results

The models were evaluated based on Mean Average Precision (MAP) as the official metric.

3.1. Task A: Multilingual Job Title Matching

We submitted five systems for Task A, varying the base model and fine-tuning techniques. The GTE-Qwen2-7B model fine-tuned with LoRA and CachedGIST achieved the best overall performance across English, Spanish, and German with an Avg.MAP of 0.52. The results are summarized in Tables 1, 2, and 3.

3.1.1. Training Procedure

All models were fine-tuned using the `sentence-transformers` library with a batch size of 64 and the AdamW optimizer. We used the CachedGISTEmbedLoss to align model outputs with embeddings generated from the `all-MiniLM-L12-v2` guide model. The training ran for 1–3 epochs with early stopping based on MAP scores on a 10% validation split.

For the GTE-Qwen2-7B model, we applied LoRA fine-tuning with a rank of 8 and trained only adapter layers to reduce GPU memory requirements. For the distilled version of `gte-multilingual-base`, we used a teacher–student setup, training the student with MSE loss to mimic embeddings from the GISTEmbed-fine-tuned full model.

3.1.2. Inference Strategy

At inference time, both query and candidate job titles were encoded using the fine-tuned model, and cosine similarity was computed to rank candidate titles. We used the same multilingual model per base encoder across all supported languages, without applying any task-specific heuristics. For Chinese job titles, LLM-generated translations of ESCO data ensured consistent structure and terminology with the English training examples.

3.1.3. Motivation for Cross-Lingual Setup and Translation Strategy

A major challenge in Task A was the lack of labeled training data for Chinese. To overcome this, we translated English job titles from ESCO into Chinese using the Qwen2.5 32B Instruct model. The motivation was to create aligned examples that preserved semantic structure while leveraging a powerful LLM's cross-lingual generation capabilities. By training a single multilingual model for all languages (en, es, de, zh), we aimed to ensure consistent semantic space alignment and reduce the complexity of maintaining separate models.

This design choice enabled the model to learn language-agnostic representations of job titles, facilitating strong cross-lingual performance as reflected in the MAP scores.

Table 1
Task A Monolingual Job Title Matching Results (MAP). Avg.MAP is over en, es, de.

Base Model	Avg.MAP	MAP(en-en)	MAP(es-es)	MAP(de-de)
GTE-Qwen2-7B (LoRA)	0.52	0.563	0.507	0.476
BGE-M3	0.48	0.496	0.479	0.458
GTE-multi-base	0.48	0.520	0.462	0.449
BGE-small-en	0.41	0.508	0.387	0.340
GTE-multi-base (Distill)	0.41	0.441	0.398	0.380

Table 2
Task A Cross-Lingual Job Title Matching Results (MAP).

Base Model	MAP(en-es)	MAP(en-de)
GTE-Qwen2-7B (LoRA)	0.525	0.504
BGE-M3	0.452	0.441
GTE-multi-base	0.453	0.449
BGE-small-en	0.325	0.307
GTE-multi-base (Distill)	0.363	0.353

Table 3
Task A Chinese Job Title Matching Results (MAP).

Base Model	MAP(zh-zh)	MAP(en-zh)
GTE-Qwen2-7B (LoRA)	0.516	0.524
BGE-M3	0.427	0.439
GTE-multi-base	0.447	0.476
BGE-small-en	0.142	0.109
GTE-multi-base (Distill)	0.448	0.396

3.2. Task B: Job Title-Based Skill Prediction

For Task B, we submitted three systems, leveraging LLM-augmented data. The results are shown in Table 4.

The GTE-Qwen2-7B model fine-tuned with LoRA on the mixed augmented data (descriptions and synonym lists) yielded the highest MAP of 0.36. This suggests that the combination of a large instruction-tuned base model, LoRA, and rich augmented input was most effective. The BGE-Large model trained solely on the augmented descriptions performed competitively (MAP 0.34). When BGE-Large was trained on the mixed augmented data, the performance was slightly lower (MAP 0.33). This might indicate that for BGE-Large, the simpler augmented descriptions were more directly beneficial, or

Table 4
Results for TalentCLEF 2025 Task B: Job Title-Based Skill Prediction.

Submission Tag Suffix	Base Model	Training Data Strategy	MAP
GTE-7B-Aug-Mix	GTE-Qwen2-7B (LoRA)	Augmented Mixed	0.36
BGE-L-Aug	BGE-Large-en-v1.5	Augmented Descriptions Only	0.34
BGE-L-Aug-Mix	BGE-Large-en-v1.5	Augmented Mixed	0.33

that the mixed data format required further hyperparameter optimization. The use of LLM-generated descriptions proved crucial for providing rich textual context.

3.2.1. Training Procedure

All models were fine-tuned using the `sentence-transformers` library with a batch size of 64, using the AdamW optimizer and a cosine learning rate schedule. We employed early stopping based on validation MAP. For GISTEmbed-based training, the guide model was set to `all-MiniLM-L12-v2`, and the CachedGISTEmbedLoss was used to align the student model's embeddings with cached guide embeddings.

For the GTE-Qwen2-7B model, we used Low-Rank Adaptation (LoRA) with a rank of 8 and bias training enabled, to efficiently fine-tune the model without updating all parameters. Fine-tuning was performed for 1–3 epochs depending on convergence behavior, monitored using a 10% validation split from the training set.

3.2.2. Inference Strategy

During inference, we used the same augmentation templates as during training. Each job title query was transformed into a prompt-generated description (optionally combined with a list of aliases), and similarly for each skill. These texts were encoded into embeddings using the fine-tuned model, and cosine similarity was computed between the job title and all skills. The top-N most similar skills were returned as the model's output for ranking.

In mixed-format submissions, we used newline-separated concatenations of generated descriptions and alias lists. This format provided the model with richer and more consistent context and led to improved generalization.

3.2.3. Motivation Behind Data Augmentation

The core motivation for data augmentation was to enrich the semantic content of both job titles and skills, which are otherwise short and ambiguous. By prompting the Qwen2.5 32B Instruct model to generate compact but expressive descriptions, we aimed to reduce lexical sparsity and improve the model's ability to match based on conceptual similarity.

Additionally, including known aliases in the input helped align representations across synonymous phrases. The combination of these strategies allowed the model to learn more generalizable embeddings, making it better suited for real-world applications where job titles and skill names vary significantly in wording.

4. Conclusion

In our participation in TalentCLEF 2025, we explored the efficacy of fine-tuning sentence embedding models using GISTEmbed for both multilingual job title matching (Task A) and job title-based skill prediction (Task B). For Task A, larger models like GTE-Qwen2-7B, fine-tuned with LoRA, demonstrated superior performance, particularly when augmented with LLM-translated data for low-resource languages like Chinese. For Task B, data augmentation via LLM-generated job and skill descriptions

was a key strategy. The GTE-Qwen2-7B (LoRA) model trained on mixed augmented data (descriptions and synonyms) achieved the best results, underscoring the value of rich, contextualized training inputs. Our experiments highlight the potential of combining advanced fine-tuning techniques like GISTEmbed and LoRA with LLM-driven data augmentation for complex semantic matching tasks in the HR domain.

Declaration on Generative AI

During the preparation of this work, the author(s) used Qwen2.5 32B Instruct in order to: Generate training data through textual descriptions for jobs and skills (Task B), and Translate English job titles to Chinese for training data augmentation (Task A). 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.

References

[1] L. Gasco, H. Fabregat, L. García-Sardiña, P. Estrella, D. Deniz, A. Rodrigo, R. Zbib, Overview of the TalentCLEF 2025 Shared Task: Skill and Job Title Intelligence for Human Capital Management, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2025.

[2] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: https://arxiv.org/abs/1908.10084.

[3] A. V. Solatorio, GISTEmbed: Guided in-sample selection of training negatives for text embedding fine-tuning, 2024. URL: https://arxiv.org/abs/2402.16829. arXiv:2402.16829.

[4] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, C-Pack: Packaged resources to advance general chinese embedding, 2023. arXiv:2309.07597.

[5] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024. arXiv:2402.03216.

[6] X. Zhang, Y. Zhang, D. Long, W. Xie, Z. Dai, J. Tang, H. Lin, B. Yang, P. Xie, F. Huang, et al., mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, 2024, pp. 1393–1412.

[7] Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie, M. Zhang, Towards general text embeddings with multi-stage contrastive learning, arXiv preprint arXiv:2308.03281 (2023).

[8] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2021. URL: https://arxiv.org/abs/2106.09685.